Optimization Methods for Heterogeneous Wireless Communication Networks: Planning, Configuration and Operation

Dem Fachbereich 18
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.)
vorgelegte Dissertation

von
M.Sc. Florian Bahlke
geboren am 17.08.1987 in Offenbach am Main

Referent:	Prof. Dr.-Ing. Marius Pesavento
Korreferent:	Prof. Dr.-Ing. Eduard A. Jorswieck
Tag der Einreichung:	29.10.2018
Tag der mündlichen Prüfung:	30.01.2019

D 17
Darmstädter Dissertation
2019

Bahlke, Florian: *Optimization Methods for Heterogeneous Wireless Communication Networks: Planning, Configuration and Operation*
Darmstadt, Technische Universität Darmstadt
Jahr der Veröffentlichung der Dissertation auf TUprints: 2019
Tag der mündlichen Prüfung: 30.01.2019

Acknowledgments

I wish to sincerely thank Prof. Dr.-Ing. Marius Pesavento for his academic guidance and support. His encouragement, friendship and the knowledge he shared with me made this work possible.

I would like to thank Prof. Dr.-Ing. Eduard M. Jorswieck, Prof. Dr.-Ing. Klaus Hofmann, Prof. Dr.-Ing. Anja Klein and Prof. Dr. mont. Mario Kupnik for their work as doctoral examiners.

The Communication Systems Group has provided me with very valuable discussions and technical lessons to improve my skills as a researcher over the last five years. I want to thank Nils Bornhorst, Yong Cheng, Dana Ciochina, Ganapati Hegde, Minh Trinh Hoang, Gerta Kushe, Tianyi Liu, Ying Liu, Fabio Nikolay, Pouyan Parvazi, Oscar Ramos, Christian Steffens, Wassim Suleiman, Dima Taleb, Xin Wen, Yang Yang and Xin Zhang for their friendship and advice. I especially want to thank Marlis Gorecki for her help in administrative matters and in the organization of teaching activities.

During my studies at TU Darmstadt, I enjoyed the company of many other students and researchers. I thank my friends and colleagues from the Communications Engineering Lab and the Signal Processing Group. My master thesis in collaboration with the Sensor Data Fusion Department at Fraunhofer FKIE has contributed greatly to my decision to pursue a doctoral degree. I want to thank my former colleagues at FKIE, especially Dr. rer.nat. Ulrich Nickel and Dr.-Ing. Reda Zemmari.

Finally, I want to express my deepest gratitude to my parents Martina and Wolfgang, my brother Philipp, my wife Eva and my daughter Mina for their continuous love, encouragement and support.

Kurzfassung

Die vierte Generation der Mobilkommunikationsnetze hat flächendeckende Verbreitung erreicht, und die kommende fünfte Generation (5G) bildet einen signifikanten Anteil der aktuellen Forschung. 5G Netzwerke sind darauf ausgelegt, in mehreren Aspekten höhere Leistung zu erreichen, und neuartige Services zu unterstützen. Je nach Anwendungsgebiet sind hierbei eine hohe Datenrate, geringe Latenz, hohe Zuverlässigkeit oder die Unterstützung einer sehr hohen Anzahl verbundener Geräte nötig. Da die erreichte Datenrate einer einfachen Punkt-zu-Punkt Verbindung bereits nahe an ihrem theoretischen Optimum liegt, müssen in 5G mehr Ressourcen aufgewendet werden um eine weitere Leistungssteigerung des Netzwerks zu erreichen. Mögliche Technologien für zukünftige Mobilkommunikationsnetze sind unter anderem die Nutzung von sehr großen Antennenarrays mit hunderten Antennenelementen oder eine Erweiterung des verwendeten Frequenzbandes in den Millimeterwellenbereich. Diese und andere Technologien verlangen signifikante Modifikationen der Netzwerkarchitektur, und damit hohe Investitionen des Netzwerkbetreibers. Eine bereits etablierte Technologie um die Leistungsfähigkeit eines Mobilkommunikationsnetzes zu erhöhen ist eine räumliche Verdichtung der Mobilfunkzellen. Dies wird erreicht indem die existierenden Zellen mit hoher Sendeleistung durch eine größere Zahl kleiner Zellen unterstützt werden, was in einem sogenannten "Heterogenen Netzwerk" (HetNet) resultiert. Dieser Ansatz erweitert die bereits existierende Architektur des Netzes und unterstützt die beschriebenen weiterführenden Technologien, welche komplexere Hardware benötigen. Heterogene Netze sind daher eine gute Übergangstechnologie für 5G und zukünftige Generationen von Mobilkommunikationsnetzen.
Die signifikanteste Herausforderung von HetNets ist dass die Verdichtung des Netzwerks für dessen Leistungsfähigkeit nur bis zu einem bestimmten Level förderlich ist. Die erreichten Datenraten sind begrenzt durch die räumlich sehr nahen benachbarten Zellen, und der ökonomische Betrieb des Netzwerks wird eingeschränkt durch den hohen Energieverbrauch und Hardwarekosten, die durch eine große Anzahl an Zellen entstehen. Diese Dissertation behandelt die Herausforderung, durch eine Verdichtung des Netzwerks zuverlässige Leistungssteigerung zu erzielen und gleichzeitig die Servicequalität und den ökonomischen Betrieb sicherzustellen.
Dieses grundlegende Problem wird auf mehreren Ebenen adressiert, die sich unterscheiden im Bezug auf den Zeithorizont in dem Maßnahmen zur Netzwerkoptimierung eingeleitet, die nötigen Informationen gesammelt, und die Optimierungen durchgeführt werden. Diese Zeithorizonte werden unterschieden in die Phasen der Planung, Konfiguration und Operation. Optimierungsverfahren für die Energie- und Ressourceneffizienz des Netzwerks werden hauptsächlich entwickelt für die Konfigurationsphase. Da ein

Netzwerk mit gleichmäßiger Lastverteilung als Basis für weitere Optimierungen dient, werden für die Planungs- und Operationsphase Verfahren entwickelt um diese zu erreichen und dauerhaft sicherzustellen.
Für die Planungsphase werden die Standorte neuer Zellen in einem existierenden Netzwerk optimiert, und die Aktivitätsphasen der Zellen geplant anhand der zu erwartenden Auslastung. Es wird gezeigt, dass eine gemeinsame Optimierung der Standorte mehrerer Zellen einer konsekutiven Aufstellung im Bezug auf die Lastverteilung des HetNets überlegen ist. Der Zeitplan für die Zellaktivität und die Länge der jeweiligen Zeitphasen werden gemeinsam optimiert. Durch dieses, aus der Verfahrenstechnik übernommene Konzept, erreicht die Planung der Aktivitätsphasen der Zellen die beste Lastverteilung. Simulationsergebnisse zeigen dass die Auslastung von überladenen Zellen effektiv verringert werden kann durch eine Optimierung der Aufstellungsorte und der Aktivität von Zellen.
Der Betrieb des Netzwerkes mit hoher Ressourceneffizienz und unter Sicherstellung der Servicequalität wird erreicht durch eine Optimierung in der Konfigurationsphase. Es wird ein Optimierungsproblem entwickelt um den Ressourcenverbrauchs des Netzwerks zu optimieren mittels mehrerer Subnetze, die orthogonal zueinander mit unterschiedlichen Ressourcen operieren. Für dieses Problem, welches für größere Netzwerke sehr hohe Komplexität aufweist, wird eine lineare innere Approximation gebildet, welche fast optimale Ressourceneffizienz erreicht. Die Interferenzen werden während der Optimierung dynamisch modelliert, wodurch im Vergleich zu gängigen Verfahren die Auslastung von Zellen genauer approximiert werden kann.
Um den höheren Energieverbrauch, welcher durch ein dichteres Netzwerk entsteht, zu verringern, wird die Sendeleistung und die Aktivität der Zellen im Netzwerk gleichzeitig optimiert. Für das sich ergebende Optimierungsproblem wird eine vereinfachte innere Approximation gebildet. Mehrere Verfahren zur Optimierung des Energieverbrauchs werden in einem simulierten HetNet getestet. Die entwickelte Methode erreicht einen niedrigeren Energieverbrauch als gängige, heuristische Verfahren, und findet in schwierigen Szenarien mit höherer Wahrscheinlichkeit eine Konfiguration für das Netzwerk, die alle Bedingungen an die Servicequalität erfüllt.
Zuletzt wird das Problem adressiert, eine ausgeglichene Lastverteilung im Netzwerk während der Operationsphase zu erhalten. Ein Verfahren basierend auf einer Mehrklassen-Stützvektormethode wird genutzt um das Lastverteilungsproblem dezentral zu lösen. Etablierte Methoden basieren häufig auf umfangreicher Kommunikation zwischen Zellen um Optimierungsprobleme zentral zu lösen. Das entwickelte dezentrale Verfahren erreicht eine fast optimale Lastverteilung obwohl die durchgeführten Optimierungen von den Mobilfunkzellen und Nutzern nur mit lokal verfügbaren Informationen durchgeführt werden.

Abstract

With the fourth generation of wireless radio communication networks reaching maturity, the upcoming fifth generation (5G) is a major subject of current research. 5G networks are designed to achieve a multitude of performance gains and the ability to provide services dedicated to various application scenarios. These applications include those that require increased network throughput, low latency, high reliability and support for a very high number of connected devices. Since the achieved throughput on a single point-to-point transmission is already close to the theoretical optimum, more efforts need to be invested to enable further performance gains in 5G. Technology candidates for future wireless networks include using very large antenna arrays with hundreds of antenna elements or expanding the bandwidth used for transmission to the millimeter-wave spectrum. Both these and other envisioned approaches require significant changes to the network architecture and a high economic commitment from the network operator. An already well established technology for expanding the throughput of a wireless communication network is a densification of the cellular layout. This is achieved by supplementing the existing, usually high-power, macro cells with a larger number of low-power small cells, resulting in a so-called heterogeneous network (HetNet). This approach builds upon the existing network infrastructure and has been shown to support the aforementioned technologies requiring more sophisticated hardware. Network densification using small cells can therefore be considered a suitable bridging technology to path the way for 5G and subsequent generations of mobile communication networks.

The most significant challenge associated with HetNets is that the densification is only beneficial for the overall network performance up to a certain density, and can be harmful beyond that point. The network throughput is limited by the additional interferences caused by the close proximity of cells, and the economic operability of the network is limited by the vastly increased energy consumption and hardware cost associated with dense cell deployment. This dissertation addresses the challenge of enabling reliable performance gains through network densification while guaranteeing quality-of-service conditions and economic operability.

The proposed approach is to address the underlying problem vertically over multiple layers, which differ in the time horizon on which network optimization measures are initiated, necessary information is gathered, and an optimized solutions are found. These time horizons are classified as network planning phase, network configuration phase, and network operation phase. Optimization schemes are developed for optimizing the resource- and energy consumption that operate mostly in the network configuration phase. Since these approaches require a load-balanced network, schemes to achieve

and maintain load balancing between cells are introduced for the network planning phase and operation phase, respectively.

For the network planning phase, an approach is proposed for optimizing the locations of additional small cells in an existing wireless network architecture, and to schedule their activity phases in advance according to data demand forecasts. Optimizing the locations of multiple cells jointly is shown to be superior to deploying them one-by-one based on greedy heuristic approaches. Furthermore, the cell activity scheduling obtains the highest load balancing performance if the time-schedule and the durations of activity periods is jointly optimized, which is an approach originating from process engineering. Simulation results show that the load levels of overloaded cells can be effectively decreased in the network planning phase by choosing optimized deployment locations and cell activity periods.

Operating the network with a high resource efficiency while ensuring quality-of-service constraints is addressed using resource optimization in the network configuration phase. An optimization problem to minimize the resource consumption of the network by operating multiple separated resource slices is designed. The originally problem, which is computationally intractable for large networks, is reformulated with a linear inner approximation, that is shown to achieve close to optimal performance. The interference is approximated with a dynamic model that achieves a closer approximation of the actual cell load than the static worst-case model established in comparable state-ot-the art approaches.

In order to mitigate the increase in energy consumption associated with the increase in cell density, an energy minimization problem is proposed that jointly optimizes the transmit power and activity status of all cells in the network. An original problem formulation is designed and an inner approximation with better computational tractability is proposed. Energy consumption levels of a HetNet are simulated for multiple energy minimization approaches. The proposed method achieves lower energy consumption levels than approaches based on an exhaustive search over all cell activity configurations or heuristic power scaling. Additionally, in simulations, the likelihood of finding an energy minimized solution that satisfies quality-of-service constraints is shown to be significantly higher for the proposed approach.

Finally, the problem of maintaining load balancing while the network is in operation is addressed with a decentralized scheme based on a learning system using multi-class support vector machines. Established methods often require significant information exchange between network entities and a centralized optimization of the network to achieve load balancing. In this dissertation, a decentralized learning system is proposed that globally balance the load levels close to the optimal solution while only requiring limited local information exchange.

Contents

Chapter 1

Introduction

1.1 5G Mobile Communication Networks

Since the fourth generation (4G) of radio access technology (RAT) in wireless communication networks has reached maturity with the widespread deployment of LTE-Advanced (LTE-A), multiple technology directions for future generations have been under extensive research during recent years. The upcoming fifth (5G) and future generations are designed to support a wide variety of network topologies and services, significantly expanding the mostly homogeneous and hierarchical architectures of current networks [ABC+14, BHL+14, WHG+14, Iwa15, NGM15]. Between the years 2017 and 2020 alone, a threefold increase in worldwide mobile data traffic to over 400 exabytes per year is forecasted [Cis17]. Not only the number of mobile devices, but also the data rates required to support novel applications drive this exponential growth in data traffic. Some of these services, such as Machine-to-Machine (M2M) communications, multimedia streaming, or virtual reality (VR) applications may require an extremely low latency, very reliable connections, enhanced support for user mobility [GJ15, Fet14], or an increased data rate. This poses novel challenges in the network design process that previously followed the aim of providing uniform user experience in every connection [SAD+16, HLQ+14]. As 5G is an evolution of the mobile communication network out of necessity for higher performance and new capabilities, its function can best be envisioned by discussing the desired use-cases established in the scientific community. The Radiocommunication Sector of the International Telecomunication Union (ITU-R) has defined three broad categories of usage scenarios for 5G [itu15, ITU17, SMS+17, XMH+17], which will be used in the following to outline the envisioned services of 5G networks. Other groupings of the same services have been proposed by researchers in the industry and the academia [3GP16, OBB+14].

Enhanced Mobile Broadband (eMBB) refers to the improvement and expansion of the current mobile network. Especially the insufficiency of data rates and seamlessness of the user experience in critical conditions demonstrate some of the shortcomings of 4G standards. In a typical cellular network architecture, users located near cell edges generally experience decreased data rates, due to high interferences from neighboring cells. The user experience in 5G however should be largely unaffected by the underlying cellular architecture of the network. The optimal network is planned to serve a wide

area, but also very concentrated hotspots of users, both with a high quality of service. Typical users in this usage scenario are private and business subscribers with mobile devices, whose main concerns are good coverage and high data rates. The desired increase in throughput however cannot be accompanied by an equivalent increase in energy consumption and operation cost [CSS+14]. A corresponding increase in spectral efficiency and energy efficiency must be achieved to enable economically operable 5G networks. Additionally a high mobility of mobile devices must be supported, for example for users in high-speed trains.

Ultra-reliable and Low Latency Communications (URLCC) constitute all usage scenarios where the limiting factors are both latency and reliability, as one rarely can be achieved without the other. Typical scenarios with such quality of service constraints are communications between vehicles in transportation systems (vehicle-to-vehicle, V2V), control mechanisms for energy grids and medical applications. Most prominent is for example the vision of conducting remote surgeries using remotely controlled robotics, which could provide complex medical services to the population of remote areas of the world. For all these applications, a failure of the system could pose health risks and financial damages to the parties involved, so a reliable and responsive connection becomes the critical quality of service requirement. Similar to eMBB, a high mobility must be supported especially for V2V services.

Massive Machine Type Communications (mMTC) applies the paradigm of an Internet of Things (IoT) in the architecture of wireless communication networks. The defining characteristic of this usage scenario is a very high spatial density of often small devices. Sensor networks and applications in machine-to-machine (M2M) communications require sophisticated protocols to increase battery lifetimes of the sensor devices [TUY14]. Latency has a lower priority than for URLCC and the throughput achieved by each connected device is much lower than in eMBB. Capabilities for decentralized self-organization of sub-networks should be supported by the underlying cellular network.

Due to the described manifold of services, it becomes immediately apparent that the existing network architecture is insufficient to support the demands for 5G. The expected gains in data rate alone cannot be fulfilled by marginal improvements in the bandwidth efficiency of LTE. Therefore an expansion of the network resources is necessary, and current research on 5G networks focuses on multiple types of such network resources. Extensive research into modulation and coding schemes has already pushed the efficiency of a single point-to-point connection in LTE-A to its theoretical limit. The expansion of wireless communication network technology for 5G therefore needs to incorporate an expansion of "physical" resources such as the invested energy, number

of antennas, frequency bandwidth, or the spatial density of the cellular architecture. It is commonly acknowledged that the most promising candidate technologies for future wireless networks build upon utilizing multiple of the aforementioned resources simultaneously. Current research favors a combination of Massive MIMO, the utilization of additional frequencies in the Millimeter-Wave band, and a significant densification of the cellular network architecture. Even though this dissertation focuses on Network Densification, it is important to address the synergies between these technologies to understand why state-of-the-art research aims to utilize them jointly.

Multiple Input Multiple Output (MIMO) systems employ multiple antennas for signal transmission and reception, and enables the transmission of multiple data streams over the same radio channel by utilizing multipath propagation. MIMO has played a decisive role for the success of current LTE and Wi-Fi systems. *Massive-MIMO* (mMIMO) [LETM14] refers to multi-antenna systems that use a very large number of antennas, which is higher than the number of users being served simultaneously, and usually over 100. Space Division Multiple Access (SDMA) is then used to provide radio links to all users with the same time-frequency resources. Recently also Non-Orthogonal Multiple Access (NOMA), building upon established research in multiuser downlink beamforming, has received increased attention [DYFP14] for an application in 5G. For a very large number of antennas, the resulting channel from the antenna array to the user becomes almost deterministic, an effect that is called channel hardening. Additionally for a large ratio of the number of antennas in the array to the number of served users, the channels are all approximately orthogonal to each other. This enables the utilization of simple linear transmit precoding and receive combining schemes [NLM13]. A significant challenge currently under investigation for mMIMO is the increased hardware cost that would be caused by using one radiofrequency (RF) chain for each antenna. These RF chains contain amplifiers, phase shifters, up/down converters and analog/digital converters. State-of-the-art approaches in mMIMO aim to decrease the cost and energy consumption for these components, or to use single units of them jointly for multiple antennas [HIXR15]. Another major test for the practical feasibility of mMIMO is the coordination of antenna beams between neighboring cells, to prevent these cells from causing significant interferences to each other [NKDA18].

The most intuitive approach to increase the amount of resources available for the wireless network is to use additional radio frequency bands for transmission. As the Ultra-High Frequency Band currently used for 4G is already very crowded, research for 5G focuses on the viability of using *Millimeter Waves* (mmWave) in the Extremely High Frequency (EHF) band above 30 GHz. In the EHF range, unused spectrum is readily available [GTC+14]. The higher transmission frequencies positively affect the channel latency, which is especially suitable for URLLC applications, whereas the

wider frequency bands and increased data rates are ideal to meet the throughput demands of eMBB. As a major disadvantage, mmWaves suffer higher propagation loss from atmosphere and rainfall than UHF waves, and even higher building penetration loss [RSM+13, XMH+17]. In return, the resulting interferences for mmWave are lower than for traditionally used frequency bands. Coverage areas of mmWave base stations are therefore expected to be only a few hundred meters in diameter, and separate access points would be necessary to achieve indoor coverage [RMSS15]. In urban areas, Line-of-Sight (LOS) between transmitter and receiver has proven to be desirable but not strictly necessary, as Non-Line-of-Sight (NLOS) transmissions have been successfully tested [RSP+14]. Since mmWave is expected to be operated using mMIMO antenna arrays with advanced RF chain technology, the discussed challenges regarding hardware cost become increasingly important [RRE14].

The third resource to be utilized in 5G is the spatial density of the deployed cells in the wireless network. If this density is increased, and therefore the size of the coverage area of each individual cell is decreased, this process is commonly referred to as *Network Densification* [BLM+14]. This is usually achieved by supplementing the existing tier of high-power macro cells (MC) with an additional tier of low-power small cells (SC) to obtain a *Heterogeneous Network* (HetNet). This method of increasing the network throughput has already been established and refined for 4G with LTE-A [KBTV10]. Multiple mechanisms to balance the network loads between the cell tiers have been developed and tested [DMW+11]. The limits of such densification in HetNets are, however, a major point of concern, primarily due to the resulting interferences [AZDG16]. Overall hardware and energy costs increase with the number of additionally deployed cells, which due to the density limitations imposed by interferences can even lead to cell deployments that do not contribute to the increase in the network performance [GTM+16]. The densification of the network therefore requires sophisticated control mechanisms that decouple the increase in throughput from a corresponding increase in harmful interferences and energy consumption [CSS+14].

To assess the applicability of the previously introduced methods mMIMO, mmWave and HetNets for 5G, their synergy in a simultaneous utilization is of paramount importance. The form factor of mMIMO antenna arrays greatly benefits from an operation in the EHF band using mWaves, because their form factor is much smaller compared to arrays with the same number of antenna elements in the UHF range [RSM+13]. The combined usage of the additional SCs deployed in a densified HetNets with mMIMO or mmWave technology however is difficult to assess in terms of performance gains. When supplementing mMIMO MCs with SCs, interference coordination between these two tiers of cells types becomes critical. Significant reductions in the overall network power consumption are achievable when the interferences are managed and the user

allocation between the cell tiers is optimized [BKD13]. Coordination between mMIMO MCs and SCs can be optimized to such a degree that the network throughput performance is mostly limited by out-of-cluster interferences from cells outside of those under consideration [JMZ+14]. It can be concluded that the combined operation of mMIMO and HetNets critically depends on the coordination between the cell tiers.

The decreased size of the coverage areas for mmWave-based cells leads to an automatic network densification, which is emphasized due to the need for separate small cell access points for indoor coverage [RMSS15]. Coverage planning, specifically the locations of SCs and MCs, needs to be executed carefully for a joint operation of mmWave MCs and SCs. The limitation of network throughput due to interferences can and must be mitigated using sophisticated interference coordination schemes [AZDG16, FWL+17]. Contrary to mmWave and mMIMO however, network densification by SC deployment constitutes an expansion of the existing network, with proven hardware components. The technological commitment and financial risk of HetNets are lower than for the other two technologies, because the latter ones require the use of advanced hardware.

It can be concluded that dense HetNets enable or support other key technologies for 5G very well. Since they build upon established hardware, and the SCs supplement an existing network, HetNets are a very good "bridging" technology to achieve throughput gains while making the necessary changes in network structure for other technologies. The key challenges associated with the SC deployment planning, network configuration, and in-operation optimization form the principal part of this dissertation.

1.2 Problem Statement

It is universally acknowledged that fundamental limits exist for the densification of a wireless cellular communication network, if said densification is to be beneficial for the network throughput [AZDG16, NK17]. The primary reason for this effect is that the amount of interference present in each connection increases with the network density, which decreases the achievable signal-to-interference-plus-noise ratios (SINR) and therefore eventually limits the achievable data rates. There exist however multiple secondary reasons for the limits of wireless network density, which include the necessity for economic operability that can be violated with increasing hardware costs and energy consumption [CSS+14, HKD11]. Additionally, the number of available deployment locations for additional base stations is limited, and each base station requires a wired or wireless data backhaul that might be subject to capacity constraints [GTM+16].

For each cell in the wireless network, the ratio of its used to its available resources defines the cell load. This load should be kept as low and, between the cells, as balanced as possible to ensure that the network can satisfy quality-of-service (QOS) constraints, while retaining agility. An overloading of single cells and an underutilization of others leads to dropped connections for the former, and is an indicator of unbalanced resource distribution. If the overall load levels can be decreased, for example through interference management or resource distribution optimization [LPGdlR+11, HRTA14, HQ14], cells free up resources that can be utilized to achieve higher data rates for their connected users. There is an equivalence between the two objectives of maximizing data rates for a limited cell load, and minimizing the cell load for fixed user rates [MK10,SY12a]. Both approaches usually achieve resource efficient solutions when performing interference management or resource distribution optimization.

Under these considerations, the following question shall summarize the main research objective of this thesis with regards to HetNets and network densification:

> How can heterogeneous wireless communication networks be planned, scheduled and operated such that an increase in cell density yields an improvement in network performance, as measured by criteria such as data rates, energy consumption and resource efficiency?

The relevance of this research objective is supported by very recent assessments about the role of dense HetNets in 5G from the scientific community. The authors in [AZDG16] state about the potential limits of network densification that *"wireless network researchers and engineers should be aware of these rapidly approaching limits, and we should begin developing communication protocols customized for dense networks"*. In [NK17], the authors further emphasize the importance of developing optimization schemes for dense HetNets: *"In practice, installing more BSs is beneficial to the user performance up to a density point, after which further densification can become harmful user performance due to faster growth of interference compared to useful signal. This highlights the cardinal importance of interference mitigation, coordination among neighboring cells and local spatial scheduling."* The significance of specialized resource allocation schemes for the technologies of 5G is summarized by the authors of [GTM+16] with: *"Massive MIMO antennas and millimeter-wave communications provide enough resource space for small cell BSs. How to utilize and optimize the resource allocation for BS relaying and self-transmission is a critical problem in 5G ultra-dense cellular networks."*

The fundamental challenge of this objective is that there are three time-horizons on which network planning, configuration and operation take place, and varying performance criteria that apply in each stage. For example, the deployment planning of additional small cells takes place on a very large timescale, and therefore has to consider long-term average cell load levels as an objective rather than instantaneous data rates. On the other hand, rate maximization for a single connection takes place on a very short timescale and therefore does not depend on long-term average load levels. The very broad research objective formulated in the question above therefore needs to be divided into sub-objectives that each concern a specific time horizon of the network.

For the network planning and configuration phase, the following objective applies:

> Objective 1: The wireless network architecture is designed with high spatial density of cells. Additional cells are deployed in suitable locations, to decrease the load levels of existing cells. The activity of the cells is scheduled such that load levels between all cells in the network are balanced. Both optimization procedures should be based on long-term averages of data traffic forecasts.

This first step of planning the physical deployment of cells and testing activity configurations for different deployment solutions typically takes place over a period of multiple weeks or months, and is accompanied by an extensive planning effort [SY13, GTM+16, KBTV10]. However, after Problem 1 is solved for a given wireless network, the cellular layout of the network architecture can be considered static. Based on a network with static architecture, further steps consider the configuration and operation of a dense HetNet, where the behavior of the network is optimized to fulfill various objectives [ABC+14, GJ15, SAD+16]. Because of the diversity of these objectives, a multitude of sub-problems besides that of load balancing arise from the central research question formulated above. The two problems that are widely considered as critical for dense HetNets, resource and energy efficiency, as discussed in Sec. 1.1, are addressed in this thesis. All optimizations of the network are subject to quality-of-service (QOS) constraints such as minimum required data rates or SINR levels. These QOS-constraints might vary depending on which services (such as eMBB, URLLC or mMTC) are requested. The problem of efficient resource distribution in a dense HetNet can therefore be formulated as follows:

> Objective 2: The allocation of users to cells and distribution of time-frequency resources in the network is to be optimized such that the resource efficiency is maximized while meeting QOS-constraints.

As discussed in Sec. 1.1, the successful operation of HetNets in 5G critically depends on economic considerations such as costs for energy consumption. An increase in network density cannot lead to an proportional increase in energy consumption. The problem of minimizing this energy consumption is formulated as follows:

> Objective 3: The energy consumption of the dense HetNet is to be minimized while meeting QOS-constraints.

Both Objectives 2 and 3 consider network optimizations that take place on a shorter timescale than Objective 1, for example in a day-ahead scheduling of the network configuration. All of the first three objectives however do not consider real-time optimization of single connections. This is because a joint optimization of the network-wide energy consumption or resource efficiency of multiple cells or the entire network is difficult to realize based on instantaneous channel state information (CSI). The CSI observed for any single connection may be outdated during the time all necessary information is gathered, the optimization problem is solved, and the optimal configurations are redistributed in the network. It can however be assumed that cells can perform decentralized optimization schemes, if they have capacity left and are not close to being overloaded.

This underlines the importance of maintaining a load-balanced state in the network. The maintenance of load balancing is a fundamental requirement for meeting QOS requirements, and to give cells sufficient head space to perform other optimizations. The problem of load balancing maintenance is formulated as follows:

> Objective 4: The dense HetNet must be maintained in a load balanced state using fast and decentralized offloading schemes. These schemes must operate based on locally available information with low communication and coordination overhead.

Objectives 1 and 4 address the fundamental requirement for the dense HetNet to meet QOS requirements and allow for further network optimization. This requirement is that a load-balanced state can be created through optimized network planning (Objective 1) and maintained during the operation of the network (Objective 4). Objective 2 and 3 presuppose a load-balanced network and address the economic operability and resource efficiency.

1.3 Contributions and Thesis Overview

The detailed outline of this dissertation is as follows:

In **Chapter 2**, the system model for the heterogeneous wireless communication network and the signal model for the radio links between cells and users are introduced. Fundamental solutions for standard network optimization objectives, such as load-balancing and SINR-maximization, are provided. The model for characterizing the different time-horizons of network operation phases is discussed.

Chapter 3 summarizes methods to reformulate and solve optimization problems with both continuous and integer parameters. Adaptations of these methods to components of the mathematical model introduced in Chapter 2 are discussed. A machine-learning based classifier is designed to serve as a resource allocation scheme for the decentralized load balancing approaches.

The following Chapters 4-7 each consider subproblems of the research objective for heterogeneous wireless network optimization defined in Sec. 1.2. Each chapter provides a discussion of the state-of-the-art and contributions specific to each objective.

In **Chapter 4**, a cell deployment scheme is addressed that selects an optimized location and cell type for the densification of an existing network through SC deployment. Multiple candidate deployment locations and cells types with varying associated costs are considered. The scheduling of cell activity over a time period is discussed for cells with energy limitations. A joint optimization is designed for the cell activity schedule and the duration of time-slots on which the resulting schedule is applied. This joint optimization significantly improves upon the state-of-the-art solution of optimizing the system with fixed time-slot durations. The proposed solutions for both the deployment and configuration problem outperform greedy and heuristic approaches, effectively addressing Objective 1 as defined in Sec. 1.2.
This chapter is based on the following publications:

- Bahlke, F.; Ramos-Cantor, O.D.; Pesavento, M.: *Budget Constrained Small Cell Deployment Planning for Heterogeneous LTE Networks*, Proceedings of the 16th IEEE Workshop on Signal Processing Advances in Wireless Communications (IEEE SPAWC), June 2015, pp. 1-5
- Bahlke, F.; Yang, J.; Pesavento, M.: *Activity Scheduling for Energy Harvesting Small Cells in 5G Wireless Communication Networks*, accepted for publication in the Proceedings of the 29th IEEE Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC 2018), September 2018

In **Chapter 5**, a configuration scheme for resource allocation in dense HetNets with heterogeneous service requirements is considered. The proposed method maximizes the resource efficiency subject to QOS-constraints by joint optimization of the dimensioning and allocation of multiple resource pools and the allocation of users to cells. The adaptive interference model introduced in this scheme shows significant performance gains compared to established state-of-the-art methods that utilize a static interference model. In this chapter, Objective 2 defined in Sec. 1.2 is discussed.
This chapter is based on the following publication:

- Bahlke, F.; Ramos-Cantor, O.D.; Henneberger, S.; Pesavento, M.: *Optimized Cell Planning for Network Slicing in Heterogeneous Wireless Communication Networks*, IEEE Communication Letters 2018, Vol. 22 (8), pp. 1676-1679

In **Chapter 6**, an energy minimization scheme for dense HetNets with joint optimization of cell transmit powers, on-off status and user allocation is considered. A inner linear approximation of the originally intractable optimization problem is derived. The reformulated problem has decreased computational complexity and enables a network operation with lower energy consumption levels than existing heuristic approaches, which provides an answer to Objective 3 as defined in Sec. 1.2.
This chapter is based on the following publication:

- Bahlke, F.; Pesavento, M.: *Energy Consumption Optimization in Mobile Communication Networks*, submitted for journal publication (preprint: https://arxiv.org/abs/1807.02651)

In **Chapter 7**, two approaches to achieve decentralized load balancing as defined by Objective 4 in Sec. 1.2 are considered. State-of-the-art approaches to user allocation and cell range expansion for load balancing require significant coordination overhead to obtain a load balanced network configuration. The two designed approaches perform user-side and cell-side decentralized load balancing using a learning-based allocation scheme that operates with information that only needs to be available locally. Simulation results show that both schemes, while operating in a decentralized manner, achieve performance close to the globally optimal load-balancing solution.
This chapter is based on the following publications:

- Bahlke, F.; Pesavento, M.: *Decentralized Load Balancing in Mobile Communication Networks*, Proceedings of the 25th IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP 2018), April 2018, pp. 3564-3568

- Bahlke, F.; Pesavento, M.: *Optimized Small Cell Range Expansion in Mobile Communication Networks Using Multi-Class Support Vector Machines*, accepted for publication in the Proceedings of the 29th European Signal Processing Conference (EUSIPCO 2018), September 2018

A final assessment and discussion of future work is provided in **Chapter 8**.

Chapter 2

System Model

2.1 Introduction

In this chapter, a model for the downlink transmissions in a heterogeneous wireless communication network is defined that serves as a mathematical framework for network optimization. The signal model for single point-to-point transmissions in the wireless communication network is given in Sec. 2.2, followed by the definition of the SINR and the data rate. As each transmission from cell to a mobile node may only use a fraction of its available resources, metrics for transmission induced load (to the cell) and the total cell load level are derived to characterize the state of the network. In Sec. 2.3, common approaches to affect the cell load by allocating mobile nodes to different cells are discussed. The first approach is to minimize the maximum load level among all cells in the network, commonly referred to as "load balancing". The second approach aims at maximizing the SINR, and accordingly minimizing the induced load, of every single connection. The wireless communication network is optimized based on network parameters which can be adjusted on varying timeframes. Sec. 2.4 concludes with an overview of the time horizons of network planning, configuration and optimization. An overview of the network optimization methods introduced in this thesis, their objectives, and the timescale on which they are applied, is also provided.

2.2 Heterogeneous Wireless Networks

A wireless communication network is considered with K cells and the set of all cells being $\mathcal{C} = \{1, \ldots, K\}$. The subsets $\mathcal{C}^{\mathrm{MC}} \subset \mathcal{C}$ and $\mathcal{C}^{\mathrm{SC}} \subset \mathcal{C}$ with $\mathcal{C} = \mathcal{C}^{\mathrm{SC}} \cup \mathcal{C}^{\mathrm{MC}}, \mathcal{C}^{\mathrm{SC}} \cap \mathcal{C}^{\mathrm{MC}} = \emptyset$ indicate macro cells (MC) and small cells (SC), respectively. The network area under consideration contains M so-called "demand points" (DP), with the set of all DPs $\mathcal{M} = \{1, \ldots, M\}$. DP $m \in \mathcal{M}$ exhibits the data rate demand d_m in bits per second, which may represent the demand of single mobile users or aggregated data demand of multiple users in a hotspot. The attenuation factor of a single-input-single-output radio link between cell $k \in \mathcal{C}$ and DP $m \in \mathcal{M}$ is determined as

$$g_{km} = g_{km}^{\mathrm{ABS}} \mathbb{E}\left(\left|h_{km}^{\mathrm{CH}}\right|^2\right) g_{km}^{\mathrm{ADP}} g_{km}^{\mathrm{PROC}} \tag{2.1}$$

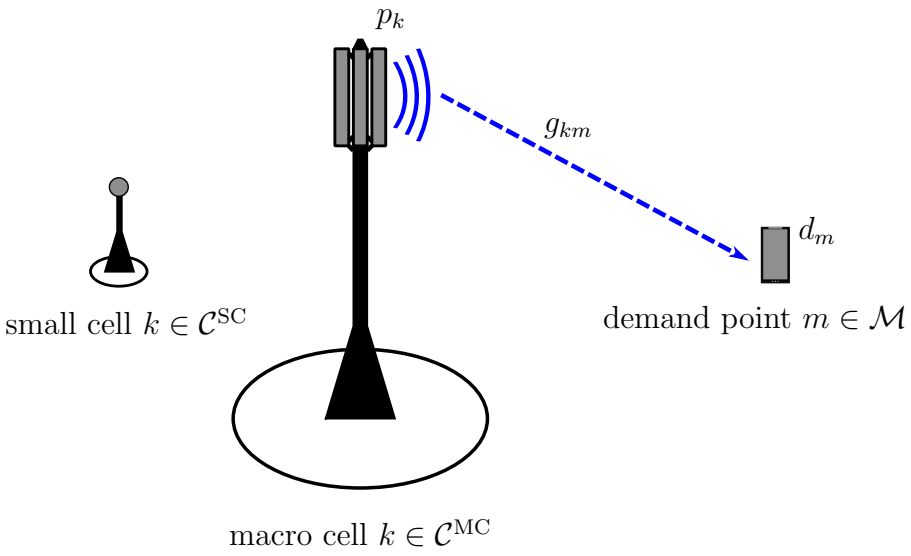

Figure 2.1. Illustration of a heterogeneous wireless network.

where g_{km}^{ABS} and g_{km}^{ADP} denote the antenna gains of the cell base station antenna and the DP antenna, respectively. The parameter g_{km}^{PROC} denotes the processing gain achieved at the receiver by multipath processing schemes such as Maximum Ratio Combining (MRC) or Zero Forcing (ZF) [Gol04, TV05]. The factor $\mathbb{E}\left(\left|h_{km}^{\mathrm{CH}}\right|^2\right)$ denotes the expected magnitude of the path attenuation. In the following, the real-valued scalar parameter $g_{km}^{\mathrm{PATH}} = \mathbb{E}\left(\left|h_{km}^{\mathrm{CH}}\right|^2\right)$ denotes the large-scale path attenuation factor caused by propagation loss and shadow fading.
The SINR of cell k serving DP m can be computed as

$$\gamma_{km} = \frac{p_k g_{km}}{\sum_{j \in \{\mathcal{C} \setminus \{k\}\}} p_j g_{jm} + \sigma^2} \tag{2.2}$$

where p_k is the transmit power of cell k and σ^2 is the power of additive white Gaussian noise, which is assumed to be identical for all DPs. The formulation $\mathcal{C} \setminus \{k\}$ refers to the set $\mathcal{C}$ without the element k. The SINR definition in (2.2) represents an orthogonal frequency-division multiple access (OFDMA) system commonly used in LTE and WLAN standards [Cim85, WCLM99, MK10]. The network is assumed to operate with full frequency reuse between cells, i.e. all cells are utilizing the same time-frequency resources. The maximum transmission rate achievable by cell k serving DP m is determined as ([MNK+07, SY12a])

$$R_{km}\left(\gamma_{km}\right) = \eta_{km}^{\mathrm{BW}} W \log_2\left(1 + \gamma_{km}\right) \tag{2.3}$$

where W is the total system bandwidth in Hz and η_{km}^{BW} is the bandwidth efficiency of the used modulation and coding scheme.
To satisfy the data demands of DP m, cell k needs to utilize at least the fraction d_m/R_{km} of its available resources. Therefore the load induced by DP m to cell k is given by

$$\frac{d_m}{R_{km}} = \frac{d_m}{\eta_{km}^{\mathrm{BW}} W \log_2\left(1+\gamma_{km}\right)}. \tag{2.4}$$

For the utilization of the cell load function in optimization problems, the following important property is proposed:

Lemma 2.2.1. *The load induced by DP m to cell k is a convex and strictly decreasing function of the SINR γ_{km} for $\gamma_{km} > 0$.*

Proof. Let

$$\zeta(\gamma) = \frac{1}{\log_2\left(1+\gamma\right)}. \tag{2.5}$$

The first and second order derivatives are given as

$$\frac{\mathrm{d}\zeta(\gamma)}{\mathrm{d}\gamma} = -\frac{\log(2)}{(1+\gamma)\log^2(1+\gamma)} \tag{2.6}$$

and

$$\frac{\mathrm{d}^2\zeta(\gamma)}{\mathrm{d}\gamma^2} = \frac{\log(2)(\log(1+\gamma)+2)}{(1+\gamma)^2\log^3(1+\gamma)} \tag{2.7}$$

Hence the lemma follows from $\mathrm{d}\zeta(\gamma)/\mathrm{d}\gamma < 0\ \forall\gamma > 0$ and $\mathrm{d}^2\zeta(\gamma)/\mathrm{d}\gamma^2 > 0\ \forall\gamma > 0$. □

To indicate the allocation of DPs to cells the binary matrix $\boldsymbol{A} \in \{0,1\}^{K\times M}$ with the matrix elements

$$A_{km} = \begin{cases} 1 \text{ if DP } m \text{ is allocated to cell } k \\ 0 \text{ otherwise} \end{cases} \tag{2.8}$$

is introduced. To satisfy the data demands of DP m, cell k needs to utilize at least the fraction d_m/R_{km} of its available resources [SY12a, MK10]. In the following, it shall be assumed that due to the used modulation- and coding scheme in the radio link under investigation, a maximum SINR level γ^{MAX} exists for which the highest possible rate is achieved, and does not improve further for $\gamma_{km} \geq \gamma^{\mathrm{MAX}}$. Let $\tau^{\mathrm{MIN}} = 1/\log_2\left(1+\gamma^{\mathrm{MAX}}\right)$ and

$$\zeta^{+}_{\tau^{\mathrm{MIN}}}(\gamma) = \max\left\{1/\log_2\left(1+\gamma\right), \tau^{\mathrm{MIN}}\right\}. \tag{2.9}$$

Therefore, the sum load of cell k that is required to serve the data demands of all its

allocated DPs (cell load) can be computed as

$$\rho_k = \sum_{m \in \mathcal{M}} A_{km} \frac{d_m}{\eta_{km}^{\mathrm{BW}} W} \zeta^{+}_{\tau^{\mathrm{MIN}}} (\gamma_{km}) \,. \tag{2.10}$$

The parameters ρ_k are the elements of the vector $\boldsymbol{\rho} \in \mathbb{R}^{K \times 1}$. For any feasible network configuration

$$0 \leq \rho_k \leq 1 \quad \forall k \tag{2.11}$$

needs to hold, as $\rho_k > 1$ would indicate that cell k is overloaded and cannot serve the data rates requested by all its allocated DPs. Note that the interference term $\sum_{j \in \{\mathcal{C} \setminus \{k\}\}} p_j g_{jm} + \sigma^2$ in the computation of the SINR Eq. (2.2) and in Eq. (2.10) can be weighted with the cell load itself [SY12a] or with an SINR-efficiency parameter [MK10] to account for the statistically lower probability that a lightly loaded cell interferes with other cells, and to consider the system's capabilities for interference mitigation. In this work, without loss of generality, the worst-case assumption that all active cells fully interfere with each other will be used. This serves as an upper bound approximation of the actual interference levels that occur while the network is in operation.

2.3 Demand Point Allocation and Load Balancing

It is assumed that a minimum SINR γ^{MIN} is required for establishing a successful wireless link between cell and DP, which is a parameter imposed by the used modulation and coding scheme. If $A_{km} = 1$ then $\gamma_{km} \geq \gamma^{\mathrm{MIN}}$ needs to hold. This can be formulated as the inequalities

$$p_k g_{km} \geq \gamma^{\mathrm{MIN}} \left(\sum_{j \in \{\mathcal{C} \setminus \{k\}\}} p_j g_{jm} + \sigma^2 \right) \quad \forall (m,k) : A_{km} = 1. \tag{2.12}$$

To avid overloaded cells in the network at all cost, a suitable optimization approach preventing such scenarios is to minimize the maximum load of any cell in the network. In the following this is referred to as "load balancing". With the continuous upper bound on the load levels Π and the allocation matrix $\boldsymbol{A}$, the following mixed integer linear optimization problem (MILP) is designed to optimize the allocation of DPs to

cells such that load balancing is achieved:

$$\begin{aligned}
\underset{\Pi, \boldsymbol{A}}{\text{minimize}} \quad & \Pi && \text{(2.13a)}\\
\text{subject to} \quad & \Pi \geq \sum_{m=1}^{M} A_{km} \frac{d_m}{\eta_{km}^{\mathrm{BW}} W} \zeta_{\tau^{\mathrm{MIN}}}^{+}(\gamma_{km}) \ \forall k && \text{(2.13b)}\\
& \sum_{k=1}^{K} A_{km} = 1 \ \forall m && \text{(2.13c)}\\
& \sum_{k} A_{km} p_k g_{km} \geq \gamma^{\mathrm{MIN}} \left(\sum_{j \in \mathcal{C}} (1 - A_{jm}) p_j g_{jm} + \sigma^2 \right) \quad \forall m && \text{(2.13d)}\\
& \Pi \in \mathbb{R}_{0+} && \text{(2.13e)}\\
& A_{km} \in \{0, 1\} \ \forall k, m && \text{(2.13f)}
\end{aligned}$$

In problem (2.13), the parameter Π in Eq. (2.13b) is the maximum load of any cell that is to be minimized. Constraints (2.13c) cause each DP to be allocated to exactly one cell. The minimum SINR condition Eq. (2.13d) is a linear reformulation of (2.12).

If the allocation of DPs to cells is not being optimized for load balancing, static allocation rules can also be employed. One such rule would be to allocate each DP to the cell that provides the strongest received signal, which maximizes the SINR of each wireless link and therefore the load each DP imposes on a cell [SY12a]. To encourage offloading to specific cells, for example the typically underutilized small cells, cell range expansion can be utilized [3GP12, SY12b, YRC+13]. The total received power $p_k g_{km}$ from cell k is multiplied with a weighting factor θ_k, the so-called "bias value", and the resulting product used for the allocation decision regarding DP m. The allocation rule can be formulated as follows:

$$A_{km} = \begin{cases} 1 & \text{if } k = \arg\max_j \theta_j p_j g_{jm} \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

where $\sum_{k \in \mathcal{C}} A_{km} = 1 \ \forall m$ needs to hold, i.e. every DP is allocated to exactly one cell. If there exist two or more cells that provide exactly the same received power according to Eq. (2.14), other, for example random, allocation rules can be used between these cells. The following property of the user allocation rule (2.14) is proposed:

Lemma 2.3.1. *The user allocation rule* (2.14) *minimizes the maximum sum load* $\Pi = \sum_{k \in \mathcal{C}}$ *with the the bias factors chosen as* $\theta_k = 1 \ \forall k$*, and uniform bandwidth efficiency* $\eta_{km}^{\mathrm{BW}} = \eta^{\mathrm{BW}} \forall k, m$.

Proof. Given (2.10) the sum load of all cells can be written as

$$\sum_{k\in\mathcal{C}} \rho_k = \sum_{k\in\mathcal{C}} \sum_{m\in\mathcal{M}} A_{km} \frac{d_m}{\eta^{\mathrm{BW}} W} \zeta^{+}_{\tau^{\mathrm{MIN}}} (\gamma_{km}) \,. \tag{2.15}$$

Due to $\sum_{k\in\mathcal{C}} A_{km} = 1$, for each DP m exactly one serving cell k is selected (by $A_{km} = 1$). To minimize the sum load of all cells, each DP m has to be served by the cell k for which it induces the lowest additional load:

$$A_{km} = 1 \quad \text{if} \quad k = \arg\min_{k^*} \left(\frac{d_m}{\eta^{\mathrm{BW}} W} \zeta^{+}_{\tau^{\mathrm{MIN}}} (\gamma_{km}) \right) \tag{2.16}$$

The function $\zeta^{+}_{\tau^{\mathrm{MIN}}} (\gamma)$ defined in (2.9) is a monotonously nonincreasing function in γ, therefore the sum cell load is minimized if the SINR γ_{km} of each individual user m is maximized:

$$\arg\min_{k} \left(\frac{d_m}{\eta^{\mathrm{BW}} W} \zeta^{+}_{\tau^{\mathrm{MIN}}} (\gamma_{km}) \right) = \arg\max_{k} \ \gamma_{km} \ \forall \ m \tag{2.17}$$

The lemma follows from $\arg\max_k \ \gamma_{km} = \arg\max_k \ p_k g_{km}$. □

The allocation rule in (2.14) can equivalently be expressed in form of the inequality

$$\sum_{k\in\mathcal{C}} A_{km} \theta_k p_k g_{km} \geq (1 - A_{jm}) \theta_j p_j g_{jm} \ \forall j, m, \tag{2.18}$$

which is used as a constraint in subsequent network optimization problems.

For each connection between cell k and DP m, the remaining cells providing the strongest second strongest interfering signals have special significance for cell load levels. These strongest interfering cells are the most significant limiting factor in achieving high data rates [MHV+12, RCBHP17a, GKN+15]. For later use, the indices of the first- and second strongest interfering cell are denoted as

$$\kappa^{\mathrm{P}}_{km} = \arg\max_{j\in\{\mathcal{C}\setminus\{k\}\}} (p_j g_{jm}) \tag{2.19}$$

and

$$\kappa^{\mathrm{S}}_{km} = \arg\max_{j\in\{\mathcal{C}\setminus\{k,\kappa^{\mathrm{P}}_{km}\}\}} (p_j g_{jm}) \tag{2.20}$$

for the connection between cell k and DP m.

2.4 Network Optimization Timescales

As the optimization of heterogeneous wireless networks incorporates multiple interdependent processes, it is essential for every network optimization scheme to first identify the timescale on which it operates [BLM+14, KBTV10, MK10]. The following three timescales shall serve as a framework for the network optimization schemes considered within this thesis:

- *Network planning phase*: This phase involves the expansion or modification of the network architecture, including base stations with baseband processors, radio-frequency frontends and antennas. Usually this phase is accompanied by extensive measurement campaigns and network simulations and takes place over the course of weeks or months. Specific examples for this step in 5G are deployment of additional small cells or a mMIMO antenna array. The deployment of additional small cells, due to the smaller transmit power and coverage area, requires a shorter planning period than a new macro cell.
- *Network configuration phase*: In the configuration stage, the architecture and the physical hardware of the network is already fixed. The resource utilization of the network components however can be optimized towards certain objectives such as load balancing, data rates or energy efficiency. Some of the network parameters such as the time-frequency resources used by each cell or the on-off status of antennas possibly cannot be changed instantaneously. Therefore, a schedule for the utilization of the resources based on data demand forecasts becomes necessary, and the optimized configuration is determined before the operation of the network, for example on the previous day.
- *Network operation phase*: The operation stage refers the network that is in-operation and all corresponding performance optimization schemes that can be applied, based on instantaneous channel feedback or short-term averages. Usually any scheme that exhibits either a high computational complexity, long inherent delays or the requirement for extensive communication- or coordination overhead is not suitable to be applied in this stage. More suitable are schemes that obtain good performance gains with limited computational effort and based on locally available information, such that they can be utilized in the range of seconds or milliseconds.

The methods for network optimization that are introduced in this thesis are each designed to be applied in one of the above stages. A summary of this classification is

Method	Chapter	Timescale	Objective
Cell deployment planning	4	planning	load balancing
Cell activity scheduling	4	configuration	load balancing
Resource planning and network slicing	5	configuration / operation	resource efficiency
Energy consumption minimization	6	configuration / operation	energy consumption
Decentralized load balancing by demand points	7	operation	load balancing
Decentralized load balancing by cells	7	operation	load balancing

Table 2.1. Method overview, timescales and objectives.

provided in Table 2.1, as well as the objectives of each optimization scheme:
Cell deployment planning aims to find the optimal locations and cell types for new cell deployments. This implicitly affects the parameters g_{km}, and the path loss between each DP and the closest cell. Deployment planning is part of the network planning stage where the network is supplemented with additional hardware. Cell activity configuration aims to find a schedule of on-off decisions for each cell for multiple consecutive time periods. This activity configuration could typically be performed in a day-ahead manner based on demand forecasts. Schemes for both cell deployment and activity scheduling with the aim to obtain a load-balanced network are introduced in Chapter 4. Resource planning introduced in Chapter 5 aims to minimize the amount of time-frequency resources required to fulfill the data demands and possibly heterogeneous service requirements of the DPs. The proposed approach is to separate the total time-frequency resources W in Eq. (2.3) into multiple independently operating resource regions, the so-called "slices". These network slices are designed based on the demands of the services they provide. This type of network optimization is suitable for optimizing a smaller network in-operation or a larger network in a resource planning schedule. The energy consumption minimization scheme introduced in Chapter 6 optimizes the on-off status and transmit power p_k of cells in order to decrease the energy consumption of the network. A real-time applicability of this scheme might be limited by startup and shutdown times of base stations. As with resource planning, the proposed approach for energy minimization is suitable for the network configuration and operation stages. Finally, the load balanced state of the network that is required for further optimization must be retained while the network is in operation. The approaches for decentralized load balancing by demand points and cells introduced in Chapter 7 are designed to operate fast and decentralized with limited coordination overhead.

Chapter 3

Methodology

3.1 Introduction

In the following an overview is provided for methods used to solve the network optimization problems that form the principal part of this thesis. Typically these problems in their original formulation are computationally intractable to solve optimally, and therefore require reformulation and approximation techniques to obtain feasible solutions and preserve scalability for larger networks. The reformulation techniques discussed in Sec. 3.2 are applied with the aim to obtain linear inner approximations or reformulations of the originally nonlinear optimization problem. A basic taxonomy of optimization problems and a motivation for aiming towards linearized problem formulations is discussed in Sec.3.2.1. Bilinear products and corresponding linear reformulation schemes are introduced in Sec. 3.2.2. Piecewise linearization of nonlinear functions, along with breakpoint selection schemes to find suitable segments for linearization, are discussed in Sec. 3.2.3. As the performance of wireless communication problems usually depends on the achievable SINR, fractional programming plays a significant role in the typical network optimization schemes. A linear reformulation technique specifically developed for fractional problems in this application scenario is discussed in Sec. 3.2.4. An introduction to Support Vector Machines (SVM), which are utilized for a fast and decentralized learning-based network load balancing scheme are introduced in Sec. 3.3. Traditionally SVMs are used for classification, but they can be adapted to solve resource allocation problems. The requirements and an outline of this SVM application are discussed in Sec. 3.3.1. An overview of training schemes for an SVM-based binary classifier are introduced in Sec. 3.3.2, which is expanded to multiclass scenarios in Sec. 3.3.3.

3.2 Mixed-Integer Programming

The network optimization problems discussed in this work are based on discrete parameters, such as binary indicators of user-cell allocations, and continuous parameters, such as the load factor of a cell. An outline of various optimization problem types and

their significance, as well as a discussion on reformulation techniques, are provided in the following.

3.2.1 Optimization Problem Taxonomy

Optimization problems containing both real and discrete parameters are classified as mixed-integer problems (MIP), whereas problems containing only integer parameters are called integer problems (IP). Both MIP and IP are NP-complete, and therefore NP-hard [Kar72]. In the specific scenario where the optimization objective function and all constraints are linear functions of all optimization variables, the problems classify as MILPs and integer linear problems (ILPs), respectively. Efficient solution algorithms for ILPs and MILPs have been continuously developed and improved since the mid of the 20th century [Dak65, Sch98, LS99].

A significant breakthrough in the theory of MILPs is that their solution can be obtained by solving a series of non-integer linear problems. This is achieved through relaxing the problem to a continuous variable space by removing the integrality constraints. The feasible solution set of the problem is then iteratively restricted with so-called "cutting planes" [Gom58], searching for solutions that are feasible for the original integer problem. If no such "integer feasible" solution is found using cutting planes, the problem is divided into sub-problems where integer parameters are fixed to different values ("branching"), and cutting planes are applied to the so obtained sub-problems. Under certain conditions it can be shown that a sub-problem cannot contain the optimal solution of the optimal problem and is therefore not further considered. This process is called "branch-and-bound", which stems from envisioning the integer problem as a decision tree. An iterative scheme combining cutting planes and branch-and-bound strategies is called "branch-and-cut", which has been a very powerful state-of-the-art approach to ILPs and MILPs in recent decades [MMWW02,CBD11]. The applications for MILPs today pervade many industries including wireless communications [ZHS10, CPP13, MCLG06]. Generic solvers for such problems are available in many programming languages [GB08, GB14, GUR, ApS17].

Contrary to MILPs, which can be reliably and efficiently solved by the aforementioned schemes, there still is no universal and established approach to mixed integer nonlinear problems (MINLPs) [BL12, KN13]. While significant advancements have been made for convex MINLPs [HBCO12, BKL$^+$13], it is universally agreed upon that nonconvex MINLPs pose a significant computational challenge where the chances of finding an optimal solution to any given problem highly depend on the problem size and structure

[FAC89, TG14]. To maintain robustness and scalability for schemes based on network optimization problems, it is therefore advisable to find an MILP that represents a linear inner approximation or a linear reformulation of the original MINLP. The problems discussed in Chapters 4, 5 and 6 are all, in their original formulation, nonconvex MINLPs. The techniques used to reformulate them are discussed in the following Secs. 3.2.2, 3.2.3 and 3.2.4.

3.2.2 Bilinear Products

Bilinear products between two optimization parameters in MILPs must be distinguished between three different types, which are integer-integer, integer-continuous and continuous-continuous products. The first two types can be recast into equivalent linear formulations using a lifting strategy, and at the cost of increased problem dimensionality [AFG04, GACD13]. These schemes will be outlined in the following.

Consider the binary parameters $b_1, b_2 \in \{0,1\}$. The product of both binary parameters is to be expressed by the auxiliary parameter $\varphi \in \{0,1\}$. The equality $b_1 b_2 = \varphi$ holds if the following inequalities are fulfilled:

$$\varphi \leq b_1 \tag{3.1a}$$
$$\varphi \leq b_2 \tag{3.1b}$$
$$\varphi \geq b_1 + b_2 - 1 \tag{3.1c}$$

A set $\mathcal{B}$ of three parameters b_1,b_2 and φ that fulfill the inequalities in (3.1), implying $b_1 b_2 = \varphi$, shall in the following be defined as

$$\mathcal{B} \coloneqq \{(b_1, b_2, \varphi) \in \{0,1\} \times \{0,1\} \times \{0,1\} : \varphi \leq b_1, \varphi \leq b_2, \varphi \geq b_1 + b_2 - 1\}. \tag{3.2}$$

Similarly, consider the binary parameter $b \in \{0,1\}$ and the real parameter $r \in \mathbb{R}$ which is bounded by $\underline{r} \leq r \leq \overline{r}$. The equality $br = \varphi$ holds if the following inequalities are fulfilled:

$$\varphi \geq r - (1-b)\overline{r} \tag{3.3a}$$
$$\varphi \leq r - (1-b)\underline{r} \tag{3.3b}$$
$$\varphi \geq \underline{r} b \tag{3.3c}$$
$$\varphi \leq \overline{r} b \tag{3.3d}$$

For $\underline{r} = 0$, the set $\mathcal{L}$ of parameters b,φ, and r with upper bound $\overline{r}$ that fulfill the inequalities in (3.3), implying $br = \varphi$, is defined as

$$\mathcal{L} \coloneqq \{(r, \overline{r}, b, \varphi) \in \mathbb{R}_{0+} \times \mathbb{R}_{0+} \times \{0,1\} \times \mathbb{R}_{0+} : \varphi \geq r - (1-b)\overline{r}, \varphi \leq r, \varphi \leq b\overline{r}\}. \quad (3.4)$$

The sets $\mathcal{B}$ and $\mathcal{L}$ are used for multiple linear reformulations of bilinear products in Chapters 4, 5 and 6. Note that if the discrete parameter in the bilinear product is an integer instead of binary, the linearization can be achieved using binary expansion [GACD13]. Let $a \in \mathbb{N}$ be a natural number with $0 < a \leq \overline{a}$, and let $L = \lfloor \log_2(\overline{a}) + 1 \rfloor$. The parameter a can be expressed as a weighted sum of binary parameters $a_l \in \{0,1\}$ with $l = 1, \ldots, L$ and

$$a = \sum_{l=1}^{L} 2^{l-1} a_l \quad (3.5)$$

for the real parameters $\varphi_l \in \mathbb{R}_{0+}$ and $\varphi = \sum_{l=1}^{L} \varphi_l$, the equality $ar = \varphi$ holds if $(r, \overline{r}, a_l, \varphi_l) \in \mathcal{L} \; \forall \; l$.

As an example based on the system model defined in Sec. 2.2, let the parameter $\Omega_{km} \in \mathbb{R}_{0+}$ define power that cell k serves DP m with, with the corresponding matrix $\mathbf{\Omega} \in \mathbb{R}_{0+}^{K \times M}$. Based on the previously defined notation, this can be expressed in a MILP as $\left(p_k, P_k^{\mathrm{MAX}}, A_{km}, \Omega_{km}\right) \in \mathcal{L} \; \forall \; k, m$, which implies $\Omega_{km} = A_{km} p_k \; \forall \; k, m$.

The product of two real parameters $r_1, r_2 \in \mathbb{R}$ that are bounded by $\underline{r_1} \leq r_1 \leq \overline{r_1}$ and $\underline{r_2} \leq r_2 \leq \overline{r_2}$ can be approximated by a set of linear inequalities using McCormick envelopes [MCB09, McC76]. The auxiliary parameter φ is used to approximate the product $r_1 r_2$ with the following inequalities:

$$\varphi \geq \underline{r_1} r_2 + r_1 \underline{r_2} - \underline{r_1}\,\underline{r_2} \quad (3.6a)$$

$$\varphi \geq \overline{r_1} r_2 + r_1 \overline{r_2} - \overline{r_1}\,\overline{r_2} \quad (3.6b)$$

$$\varphi \leq \overline{r_1} r_2 + r_1 \underline{r_2} - \overline{r_1}\,\underline{r_2} \quad (3.6c)$$

$$\varphi \leq r_1 \overline{r_2} + \underline{r_1} r_2 - \underline{r_1}\,\overline{r_2} \quad (3.6d)$$

The approximation of $\varphi = r_1 r_2$ with the above inequalities (3.6) has the critical drawback that it is neither a strict over- nor under-approximation. In the network optimization problems discussed in the following chapters, QOS constraints usually only allow an inner approximation of the original problem, i.e. every solution obtained from solving the approximated problem must be feasible for the original problem. Therefore such reformulations that lead to bilinear functions of two continuous parameters are generally avoided.

3.2.3 Piecewise Linearization

The problem of fitting a piecewise linear function to a given set of datapoints can be accomplished with linear regression and other established approaches [MB09]. If a piecewise linear function should be fitted to a given non-linear function, it may be insufficient to choose a uniform grid of discrete points on said function and then again use regression algorithms. Breakpoint selection schemes have been proposed [LT15] to find an optimized set of points on the non-linear function, where the piecewise linear segments are determined by connecting neighboring pairs of breakpoints [LCG+13]. Let $f(x)$ be a continuous function for which a piecewise linear approximation is to be found in the interval $x^{\mathrm{MIN}} \leq x \leq x^{\mathrm{MAX}}$. This objective is equivalent to finding a set $\mathcal{X}$ of breakpoints x_i^{B} with $i = 1, \ldots, I+1$ and $x_i^{\mathrm{B}} < x_{i+1}^{\mathrm{B}} \; \forall \; i \leq I$. Let

$$u_i(x) = \alpha_i x + \beta_i \tag{3.7}$$

be the linear function obtained from connecting the points $\left(x_i^{\mathrm{B}}, f\left(x_i^{\mathrm{B}}\right)\right)$ and $\left(x_{i+1}^{\mathrm{B}}, f\left(x_{i+1}^{\mathrm{B}}\right)\right)$, specifically

$$\alpha_i = \frac{f\left(x_{i+1}^{\mathrm{B}}\right) - f\left(x_i^{\mathrm{B}}\right)}{x_{i+1}^{\mathrm{B}} - x_i^{\mathrm{B}}} \tag{3.8}$$

and

$$\beta_i = f\left(x_i^{\mathrm{B}}\right) - \alpha_i x_i^{\mathrm{B}}, \tag{3.9}$$

The piecewise linearization of $f(x)$ in the interval $x_1^{\mathrm{B}} \leq x \leq x_{I+1}^{\mathrm{B}}$ shall be denoted as

$$\mathrm{Lin}_{\mathcal{X}}\left(f(x)\right) = u_i(x) \quad \text{with} \quad x_i^{\mathrm{B}} < x \leq x_{i+1}^{\mathrm{B}} \tag{3.10}$$

The problem of finding suitable breakpoints x_i^{B} can be accomplished using iterative schemes [LT15]. An analytic minimization of the number of linear functions might not be feasible, depending on the function to be linearized. Let the x-position of the maximum approximation error of a given linearization $\mathrm{Lin}_{\mathcal{X}}$ be

$$x^{\mathrm{E}} = \arg\max_x \left|\mathrm{Lin}_{\mathcal{X}}\left(f(x)\right) - f(x)\right| \tag{3.11}$$

If the approximation error should be kept below a selectable ϵ, the construction of a set of breakpoints $\mathcal{X}$ and corresponding linear functions $u_i(x)$ can be conducted as follows:

1. the set of breakpoints is initialized with the endpoints of the interval to be linearized: $\mathcal{X} = \left\{x^{\mathrm{MIN}}, \leq x^{\mathrm{MAX}}\right\}$

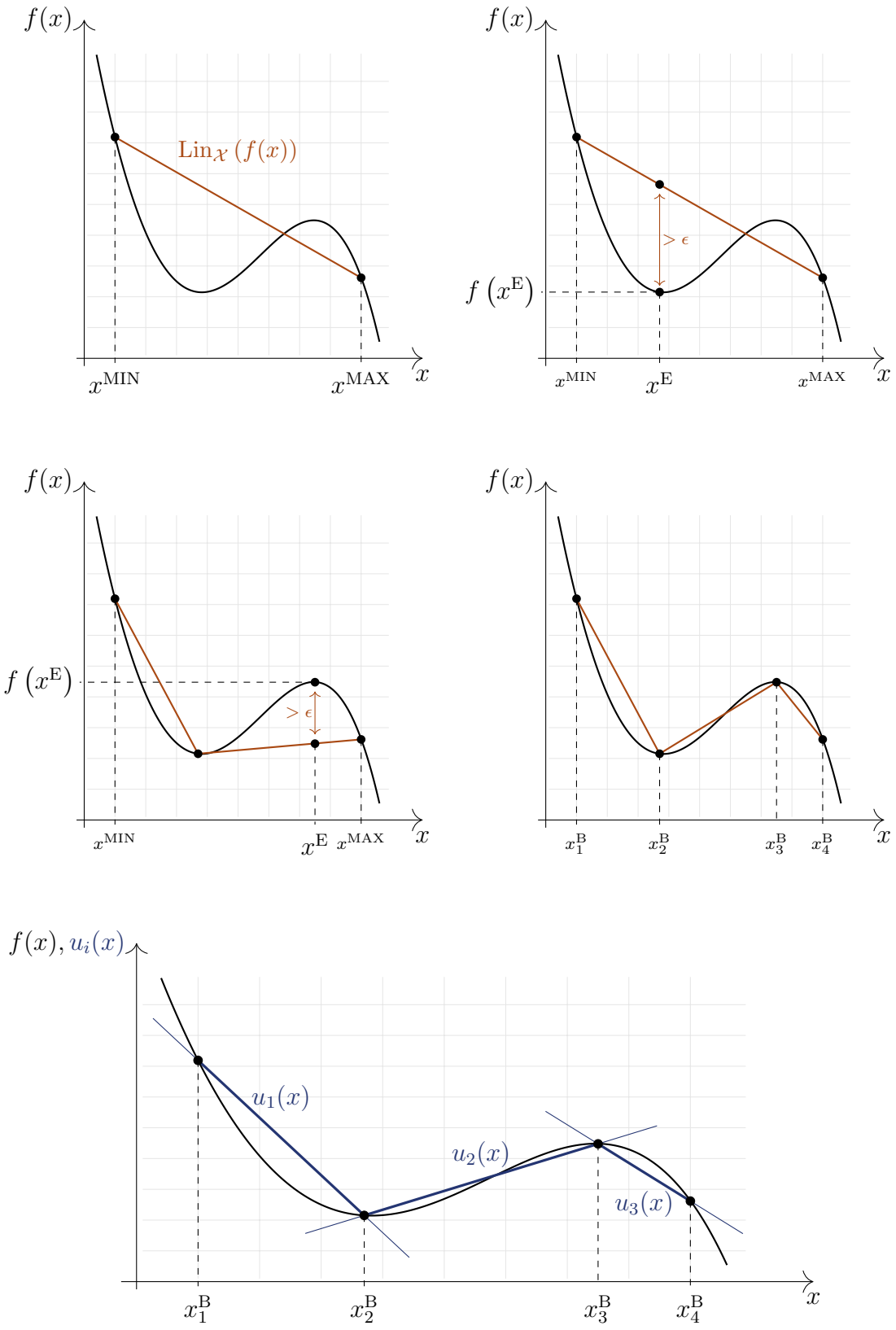

Figure 3.1. Illustration of an iterative breakpoint selection scheme for piecewise linear approximation.

2. based on $\mathcal{X}$, compute linear functions $u_i(x)$ according to Eq. (3.7)

3. compute x^{E} with Eq. (3.11)

4. if $\left|\mathrm{Lin}_{\mathcal{X}}\left(f\left(x^{\mathrm{E}}\right)\right) - f\left(x^{\mathrm{E}}\right)\right| > \epsilon$, add x^{E} to the set of breakpoints $\mathcal{X}$ and continue from step 2, otherwise return final set $\mathcal{X}$ and corresponding linear functions $u_i(x)$

The described breakpoint selection and linearization process is illustrated in Fig. 3.1. A formulation of the piecewise linearization of $f(x)$, based on linear inequalities to be utilized in a MILP, is outlined in the following. Let $\nu_i \in \{0,1\} \forall i$ be an indicator used to select the appropriate line segment, with the corresponding vector $\boldsymbol{\nu} \in \{0,1\}^{I\times 1}$. For a given x, $\mu = \mathrm{Lin}_{\mathcal{X}}\left(f(x)\right)$ can be obtained from the following optimization problem:

$$\begin{aligned}
\underset{\mu,\boldsymbol{\nu}}{\text{minimize}} \quad & \mu && \text{(3.12a)}\\
\text{subject to} \quad & \mu \geq \sum_{i=1}^{I} u_i\left(x\right)\nu_i && \text{(3.12b)}\\
& \nu_i x_i^{\mathrm{B}} \leq x \leq \nu_i x_{i+1}^{\mathrm{B}} \;\forall\; i \leq I && \text{(3.12c)}\\
& \sum_{i=1}^{I} \nu_i = 1 && \text{(3.12d)}\\
& \nu_i \in \{0,1\} \;\forall\; i \leq I && \text{(3.12e)}
\end{aligned}$$

Piecewise convexity of $f(x)$ can be exploited to decrease the number of additional parameters required in (3.12) [LCG+13]. If $f(x)$ is strictly convex, this linearization can be formulated without the segment selection parameter $\boldsymbol{\nu}$ in (3.12). For a given x, the parameter $\mu = \mathrm{Lin}_{\mathcal{X}}\left(f(x)\right)$ can be obtained from the following optimization problem:

$$\begin{aligned}
\underset{\mu}{\text{minimize}} \quad & \mu && \text{(3.13a)}\\
\text{subject to} \quad & \mu \geq u_i(x) \;\forall\; i \leq I && \text{(3.13b)}
\end{aligned}$$

As an example for this piecewise linearization, the load function $\zeta(\gamma)$ as defined in Eq. (2.5) in Sec. 2.2 is to be linearized in the interval $\gamma^{\mathrm{MIN}} \leq \gamma \leq \gamma^{\mathrm{MAX}}$. As shown in Lemma 2.2.1, $\zeta(\gamma)$ is convex and strictly decreasing in γ. The function $\zeta(\gamma)$ and corresponding $u_i(\gamma)$ that serve as a linear over-approximation are illustrated in Fig. 3.2. The approximation error between a linear function $u_i(\gamma)$ connecting two breakpoints on $\zeta(\gamma)$ and $\zeta(\gamma)$ is

$$u_i(\gamma) - \zeta(\gamma) = \alpha_i\gamma + \beta_i - \frac{1}{\log_2(1+\gamma)} \tag{3.14}$$

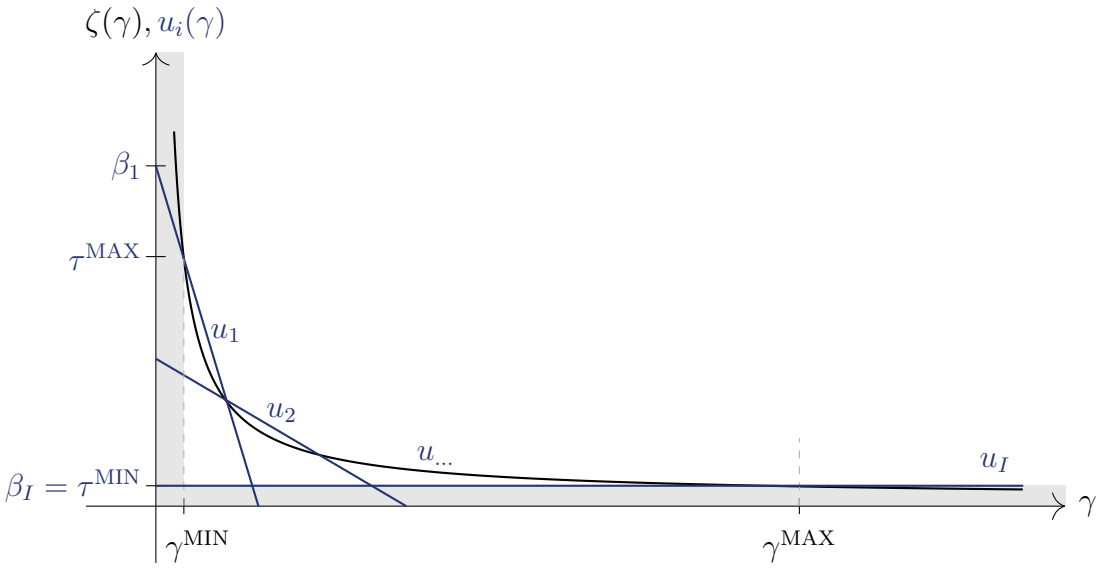

Figure 3.2. Illustration of the piecewise linear over-approximation of the cell load function $f(\gamma)$ with the linear functions $u_i(\gamma)$ in the SINR interval $\gamma^{\mathrm{MIN}} \leq \gamma \leq \gamma^{\mathrm{MAX}}$.

with the derivative

$$\frac{\mathrm{d}\left(u_i(\gamma) - \zeta(\gamma)\right)}{\mathrm{d}\gamma} = \alpha_i + \frac{\log(2)}{(\gamma+1)\log^2(\gamma+1)}. \tag{3.15}$$

In order to find the γ-position of the potential breakpoint, Eq. (3.11) is evaluated:

$$\gamma^{\mathrm{E}} = \arg\max_{\gamma} \left(u_i(\gamma) - \zeta(\gamma)\right) \tag{3.16}$$

This implies

$$\frac{\mathrm{d}\left(u_i(\gamma^{\mathrm{E}}) - \zeta(\gamma^{\mathrm{E}})\right)}{\mathrm{d}\gamma^{\mathrm{E}}} = 0 \tag{3.17}$$

which for $\alpha_i < 0$ is satisfied for

$$\gamma^{\mathrm{E}} = e^{2\mathcal{W}\left(\frac{1}{2}\sqrt{-\frac{\log(2)}{\alpha_i}}\right)}. \tag{3.18}$$

where $\mathcal{W}$ is the Lambert W-Function defined as

$$x = f^{-1}\left(xe^x\right) = \mathcal{W}\left(xe^x\right). \tag{3.19}$$

3.2.4 Fractional Bounding Discretization

Let $\boldsymbol{x}$ with elements $x_l, l \in \mathcal{F}$, $\mathcal{F} = \mathcal{F}^{\{\mathrm{N}\}} \cup \mathcal{F}^{\{\mathrm{R}\}}$, $x_l \in \mathbb{R} \forall l \in \mathcal{F}^{\{\mathrm{R}\}}$ and $x_l \in \mathbb{N} \forall l \in \mathcal{F}^{\{\mathrm{N}\}}$, i.e. the vector $\boldsymbol{x}$ contains both real and integer elements. Furthermore let $f^{\mathrm{N}}(\boldsymbol{x}) > 0$ and $f^{\mathrm{D}}(\boldsymbol{x}) > 0$ be linear functions of $\boldsymbol{x}$. A fractional MIP shall be defined as follows:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \frac{f^{\mathrm{N}}(\boldsymbol{x})}{f^{\mathrm{D}}(\boldsymbol{x})} \tag{3.20a}$$

$$\text{subject to} \quad x_l \in \mathbb{R}_{0+} \; \forall \; l \in \mathcal{F}^{\{\mathrm{R}\}}, x_l \in \mathbb{N} \; \forall \; l \in \mathcal{F}^{\{\mathrm{N}\}} \tag{3.20b}$$

Solution approaches for fractional MIPs have been proposed [YGGY13, Wu97], usually relying on variable transformations to a bilinear, and then to a MILP, using methods such as the ones discussed in Sec. 3.2.2. If the fractional term however is not in the objective, but rather appears in the constraints of a larger problem, the proposed methods might not be applicable. Additionally, the variable transformations required in [YGGY13, Wu97] cannot be applied if other constraints in the optimization problem require the original variables. Multiple iterative approaches for continuous fractional problems have been proposed with applications on wireless communications [ZJ15], but they cannot necessarily be applied if the fractional term contains integer optimization parameters. Therefore in the following a method to perform a inner approximation of the fractional MIP shall be proposed. Let $\Psi_n \in \mathbb{R}_{0+}$ with $n = 1, \ldots, N$ and let $\phi_n \in \{0, 1\}$ with the corresponding vector $\boldsymbol{\phi} \in \{0, 1\}^{N \times 1}$. Problem (3.20) is approximated with

$$\underset{\boldsymbol{x}, \boldsymbol{\phi}}{\text{minimize}} \quad \sum_{n=1}^{N} \phi_n \frac{f^{\mathrm{N}}(\boldsymbol{x})}{\Psi_n} \tag{3.21a}$$

$$\text{subject to} \quad \sum_{n=1}^{N} \phi_n \Psi_n \geq f^{\mathrm{D}}(\boldsymbol{x}) \tag{3.21b}$$

$$\sum_{n=1}^{N} \phi_n = 1 \tag{3.21c}$$

$$x_l \in \mathbb{R}_{0+} \; \forall \; l \in \mathcal{F}^{\{\mathrm{R}\}}, x_l \in \mathbb{N} \; \forall \; l \in \mathcal{F}^{\{\mathrm{N}\}} \tag{3.21d}$$

$$\phi_n \in \{0, 1\} \; \forall \; n \tag{3.21e}$$

Note that the problem (3.21) is an inner approximation of problem (3.20), i.e. every $\boldsymbol{x}$ obtained as the optimal solution of (3.21) is feasible for problem (3.21). How tight this approximation is depends on the discrete values Ψ_n and how closely they approximate the actual denominator $f^{\mathrm{D}}(\boldsymbol{x})$. Specifically, if (3.21) solves (3.20) optimally, Eq. (3.21b) is fulfilled with equality. The following situations are beneficial for the tightness of this

approximation:

- the target set of $f^{\mathrm{D}}(\boldsymbol{x})$ is integer, and can be directly represented by corresponding Ψ_n
- the fractional term is embedded as a constraint in a larger optimization problem, and only a small fraction of the target set of $f^{\mathrm{D}}(\boldsymbol{x})$ has significant effect on the overall objective
- the optimal solution very likely features a small subset of the target set of $f^{\mathrm{D}}(\boldsymbol{x})$

In Chapters 5 and 6, the proposed reformulation method for fractional terms is utilized in optimization problems that meet the aforementioned conditions.

3.3 Classifier-Based Optimization

In the following an adaptation of learning-based classifiers to the network optimization problems encountered in wireless networks is discussed. The classifier is based on support vector machines (SVM) that are extended for multi-class scenarios.

3.3.1 Allocation and Classification

Classification in the context of machine learning is the attempt to identify which class out of a given set of classes an entity belongs to. This entity usually possesses certain attributes which are either directly used as or transformed into features, based on which the assignment to the correct class should be made. In supervised learning, the classification scheme is "trained", using a dataset of entities with features and the correct classes. The trained classifier is then used with a new dataset of attributes to estimate the unknown classes.
The application of statistical learning methods in optimizing wireless communications networks is only being considered recently [JZR+17]. Wireless network operators typically face optimization problems of setting parameters or distributing resources, both usually under multiple side constraints. The following conditions are beneficial for the improvised usage of a classifier to solve the optimization problem:

- The original problem is an IP where a high number of network entities have to each choose between a limited number of options.

- If one such decision made by the classifier violates a QOS-constraint in the original problem, a fallback solution needs to be available for that entity, since hard constraints usually cannot be implemented is classifiers.

- Each entity needs to be able to extract a sufficient amount of attributes about the network state on which the decision is made. In wireless networks, this includes for example channel conditions, load levels and data rates.

These conditions are met for the ILPs discussed in Chapter 7, which are used to perform decentralized load balancing, as defined in Sec. 2.3. The decentralization of decision making is possible because the discussed features required to make the classification decision are extracted locally. Network-wide information exchange, that is necessary for the ILP-based optimization, can thus be avoided using this learning-based approach.

3.3.2 Support Vector Machines

Let $\boldsymbol{h}_{\tilde{t}}$ be a dataset of attributes with $\tilde{t} = 1, \ldots, \tilde{T}$ and let

$$\boldsymbol{H} = \left[\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{\tilde{T}}\right]^\top . \tag{3.22}$$

The class labels of the training data is given in the vector $\boldsymbol{y} = \left[y_1, \ldots, y_{\tilde{T}}\right]^\top$. During SVM training, a hyperplane $\boldsymbol{\omega}^\top \boldsymbol{h} + b = 0$ is to be found that best separates the feature datapoints into two classes. SVMs are large-margin classifiers, which means that they aim to maximize the margin $2/|\omega|$ between the hyperplane and the closest datapoints. Since usually the data cannot be separated optimally, two modifications for SVMs have been established. The first modification is soft threshold training, where the hyperplane does not have to strictly separate the two classes of datapoints. The resulting mis-classifications are discouraged during the training of the SVM. The second modification is the kernel trick, where a function $\vartheta(\boldsymbol{x}_{\tilde{t}})$ maps the attribute vector $\boldsymbol{x}_{\tilde{t}}$ onto the L-dimensional feature space. In this feature space for example polynomial combinations of the attributes are used as training features. The following optimization problem is solved to train an SVM that classifies between classes c_1 and

c_2 [Kre99, CV95]:

$$\underset{\boldsymbol{\omega}^{\{c_1c_2\}},b^{\{c_1c_2\}},\boldsymbol{\psi}^{\{c_1c_2\}}}{\text{minimize}} \quad \frac{1}{2}(\boldsymbol{\omega}^{\{c_1c_2\}})^\top\boldsymbol{\omega}^{\{c_1c_2\}} + C\sum_{\tilde{t}=1}^{\tilde{T}}\psi_{\tilde{t}}^{\{c_1c_2\}} \tag{3.23a}$$

$$\text{subject to} \quad (\boldsymbol{\omega}^{\{c_1c_2\}})^\top\vartheta(\boldsymbol{h}_{\tilde{t}}) + b^{\{c_1c_2\}} \geq 1 - \psi_{\tilde{t}}^{\{c_1c_2\}} \quad \text{if} \quad y_{\tilde{t}} = c_1 \tag{3.23b}$$

$$(\boldsymbol{\omega}^{\{c_1c_2\}})^\top\vartheta(\boldsymbol{h}_t) + b^{\{c_1c_2\}} \leq \psi_{\tilde{t}}^{\{c_1c_2\}} - 1 \quad \text{if} \quad y_{\tilde{t}} = c_2 \tag{3.23c}$$

$$\psi_{\tilde{t}}^{\{c_1c_2\}} \geq 0 \tag{3.23d}$$

$$\boldsymbol{\omega}^{\{c_1c_2\}} \in \mathbb{R}^{L\times 1}, b^{\{c_1c_2\}} \in \mathbb{R}, \boldsymbol{\psi}^{\{c_1c_2\}} \in \mathbb{R}^{L\times 1} \tag{3.23e}$$

The penalty term $C\sum_{\tilde{t}=1}^{\tilde{T}}\psi_{\tilde{t}}^{\{c_1c_2\}}$ in Eq. (3.23a), with a selectable weighting factor C, is used to discourage mis-classifications, providing the aforementioned soft-thresholding. The classifier between classes c_1 and c_2 is defined by the parameters $\boldsymbol{\omega}^{\{c_1c_2\}}$ and $b^{\{c_1c_2\}}$ obtained from solving problem (3.23). If classification has to be conducted only between these two classes, the estimated class $\hat{y}$ for a new dataset $\hat{\boldsymbol{h}}$ can be determined as follows:

$$\hat{y} = \begin{cases} c_1 & \text{if} \quad (\boldsymbol{\omega}^{\{c_1c_2\}})^\top\vartheta(\hat{\boldsymbol{h}}) + b^{\{c_1c_2\}} \geq 0 \\ c_2 & \text{if} \quad (\boldsymbol{\omega}^{\{c_1c_2\}})^\top\vartheta(\hat{\boldsymbol{h}}) + b^{\{c_1c_2\}} < 0 \end{cases} \tag{3.24}$$

Multiclass SVM training problems like (3.23) are typically solved with high computational efficiency in their Lagrange dual formulation using kernel functions [MMR+01]. This functionality is included in common machine learning software tools [CL11, MAT].

3.3.3 Multiclass Extensions

Multiple approaches exist that aim to expand the capabilities of SVM for multi-class problems [HL02, CS02]. A brief outline of the different options shall be given in the following:

- *One-Against-All*: Multiple SVMs are trained with one class being assigned the positive labels $y = 1$ and all other classes being assigned the negative labels $Y = -1$. For $i = 1, \ldots, I$ classes, the so obtained parameters $\boldsymbol{\omega}^i$ and b^i are used to estimate the class $\hat{y}$ for a new dataset $\hat{\boldsymbol{h}}$ based on the largest margin out of any of the trained SVMs:

$$\hat{y} = \arg\max_i \left((\boldsymbol{\omega}^{\{i\}})^\top\vartheta(\hat{\boldsymbol{h}}) + b^{\{i\}} \right) \tag{3.25}$$

- One-Against-One: The classification is carried out by training SVMs between all possible pairings of classes c_1 and c_2 according to (3.23). The estimated class $\hat{y}$ is chosen according to which class "wins" the most one-on-one classifications with all other classes, specifically

$$\hat{y} = \arg\max_i \left(\sum_{j=1}^{I} H\left((\boldsymbol{\omega}^{\{ij\}})^\top \vartheta(\hat{\boldsymbol{h}}) + b^{\{ij\}} \right) \right) \tag{3.26}$$

  where $H(\cdot)$ is the Heaviside step function.

- More sophisticated schemes based on decision trees have also been proposed, for example directed acyclic graph SVMs [PCST00].

Since the optimization problems encountered in wireless communication networks typically have more than two options for resource allocation or parameter settings, the extensions to multi-class SVMs are crucial to effectively solve the load balancing problems discussed in Chapter 7.

Chapter 4

Small Cell Deployment and Activity Scheduling

4.1 Introduction and Contributions

The successful operation of small cells requires an optimized planning of deployment locations [SY13], scheduling of activities, as well as allocation mechanisms for users to cells [YY17, GTM+16, KBTV10, HRTA14]. Locations have to be chosen by considering the average user demand in the serving area, the interference levels, and the possible benefit in terms of decreased load for the macro cells. Simple approaches like selecting the location solely based on the hotspot positions or the distance to macro nodes do not account for the complexity of wireless networks especially in urban areas. Each deployment is associated with specific costs that depend on the location and the type of base station. For example, it may be very expensive or not affordable to deploy a cell in certain areas, and the acquisition cost increases with enhanced capabilities, such as an increased number of transceivers or higher transmit power. In this chapter, two approaches are introduced to optimize the deployment location and small cell type selection in LTE networks, where heterogeneous distributions of the user demand and the location-dependent acquisition cost are also considered. Adaptive switching between on and off states for small cells has been proposed, against the issues of increased energy consumption and throughput-limiting interferences, and in order to utilize only those small cells that are most beneficial for the overall network performance [WWH+17, GTM+16, NH09, NK17, HKD11]. This activity scheduling becomes even more crucial when the small cells use renewable energy sources and power storage. In this work, a scheme is proposed where the optimal energy harvesting small cell activity schedule for load balancing is obtained as the solution of a mixed-integer optimization problem. For the cell activity optimization, demand forecasts are available in form of so-called network "snapshots" that capture the forecasted user data demand per area. It is assumed that the number of snapshots is very large, such that a joint optimization of the small cell activity schedule for all time periods corresponding to the snapshots is computationally intractable. Therefore, a cost function is proposed based on the changes in the demand profile between consecutive snapshots. This cost function is used to group snapshots to time-slots in a way that more time-slots per time period are used if there is a larger fluctuation in the demand profile. The obtained time-slots, each

corresponding to the time period of multiple snapshots, are then used as the timescale for the small cell activity schedule optimization.

4.1.1 State-of-the-Art

The deployment problem has been discussed before for macro base stations in 3G [ACM03] and small cells 4G networks [SY13, KMK12]. In [SY13], the authors propose an approach to small cell deployment planning that aims at load minimization by optimization of a mixed-integer nonlinear problem. The problem formulation proposed there is a MILP except for the interference term in the load computation Eq. 2.10 being weighted with the cell load of interfering cells. As discussed in Sec. 2.2, this weighting factor will not be considered in this thesis, so the problem discussed in [SY13], adapted to the system model introduced in Sec. 2.2 is a MILP. An approach for solving the deployment problem based on tabu search is proposed in [SY13] in an effort to decrease the computational complexity at the cost of potentially obtaining a sub-optimal solution. The approach proposed in [KMK12] attempts to minimize the number of additional SCs that have to be deployed to meet QoS requirements. A greedy solution approach is proposed to solve the original deployment problem sub-optimally. A selection among multiple small cell types and area-dependent acquisition costs has not been considered in the aforementioned references and the preceding literature. Activity scheduling schemes have been developed with the objective of energy minimization or rate maximization [KU16,MGRD17,LSB$^+$16,SBSLa14]. The authors in [KU16] provide a common problem formulation for the activity scheduling problem as a MILP, and a low-complexity continuous approximation. In [MGRD17] and [LSB$^+$16], learning-based solutions for similar problems have been introduced. The approaches in [KU16], [SBSLa14] and [LSB$^+$16] all focus on a minimization of the energy consumption and are unsuitable for load balancing. The learning-based approach in [MGRD17] aims to maximize the network throughput in terms of data rates. Approaches that consider load balancing with a joint optimization of the small cell activities and the timescale on which the scheduling solutions are applied however are not considered in the literature. Joint optimization of the system parameters and the time-slot durations during which individual solutions are applied is an approach well established in process engineering [SP96, FL05]. In the proposed approach, these principles are applied to a wireless communication network where a small cell activity schedule has to be optimized for a longer time period, based on many data demand forecasts.

4.1.2 Contributions and Overview

A network planning approach for the deployment of small cells in a mobile communication network is proposed that aims to deploy the best type of cell in the best location, in order to achieve load balancing for the existing network. Multiple heterogeneous parameters of the network, such as area-dependent deployment costs and different selectable small cells types, are considered in the optimization process, which expands upon the homogeneous network models used in [SY13, KMK12]. The activity of the small cells over a given time period is optimized in a scheduling scheme that performs energy management based on data demand forecasts. The length of the time periods in which the optimized scheduling is applied is jointly optimized, which has been applied before in process management [SP96, FL05], but has not been considered in cell activity scheduling for wireless networks.

The remainder of this chapter is structured as follows: Approaches for an optimization of the deployed small cell type and location are introduced in Sec. 4.2. The greedy approach to this location optimization is shown described in Sec. 4.2.1, followed by the mixed-integer programming approach in Sec. 4.2.2. For the network with deployed small cells, a energy management and activity scheduling scheme is presented in Sec. 4.3, where first the energy management is optimized in Sec. 4.3.1 and then the timescale on which the solution is applied in Sec. 4.3.2. Simulation results are presented in Sec. 4.4 and a final summary and assessment of the proposed methods is given in Sec. 4.5.

4.2 Location Optimization

Based on the system model introduced in Sec. 2.2, the established model is in the following expanded for a network where different small cell models are to be deployed in a network with multiple so-called “candidate sites” for deployment. An intuitive approach to select these candidate sites for small cell deployment is to select locations corresponding to pixels which require many resources from a cell in order to satisfy their user demands. In these pixels, either the demand is very high or the achievable SINR of the allocated base station is very low, as for example at the cell edges. Small cells deployed in the corresponding locations can assist in fulfilling the user demand in that area with a high proximity gain. Given a testing DP m, a “site suitability

indicator" can be modeled as

$$\mathrm{SSI}(m) = \sum_{m^*:||m^*-m||<R^{\mathrm{SC}}} \rho_k^2 \frac{d_{m^*}}{R_{km^*}} \tag{4.1}$$

where R^{SC} is the expected radius of a small cell coverage area and $m^* : ||m^*-m|| < R^{\mathrm{SC}}$ are all DPs m^* within this radius around the DP under investigation. To encourage offloading to overloaded cells the SSI of pixel locations that are allocated to overloaded cells is emphasized by including the weighting factor ρ_k, which is squared to highly prioritize cells with $\rho_k > 1$. In the following Secs. 4.2.1 and 4.2.2 it is assumed that, based on the highest entries obtained from the evaluation of the SSI in (4.1), a set $k \in \mathcal{C}^{\mathrm{SC}}$ of small cell candidate sites has been determined.

The dependency of the SINR on the activity and type of all interfering cells, including SCs, poses difficulties in the computation of the SINR in Eq. (2.2). Since SCs usually exhibit a much lower transmit power, and therefore smaller coverage area, than MCs, they create significantly lower interference for the DP. In order to obtain a formulation for the SINR that is computationally tractable in optimization problems for network planning and network scheduling, the interference of SCs is in this Chapter 4 neglected, and the SINR is computed as

$$\gamma_{km} = \frac{p_k g_{km}}{\sum_{j\in\{\mathcal{C}\setminus\{k,\mathcal{C}^{\mathrm{SC}}\}\}} p_j g_{jm} + \sigma^2}. \tag{4.2}$$

Furthermore, it is assumed in the following Secs. 4.2.1 and 4.2.2 that suitable candidate sites for small cells deployment have already been identified and that each $k \in \mathcal{C}^{\mathrm{SC}}$ represents one such potential deployment location, but not necessary a deployed small cell. Suitable locations for these candidate sites are, for example, the edges between the coverage areas of macro cells, areas with high data demand from DPs (hotspots) or remote areas that due to bad SINR conditions cause high load to serving macro cells. Denote as $\tilde{N}$ the number of available small cell types, indicated by $\tilde{n} = 1, \ldots, \tilde{N}$. To describe the small cell deployment configuration the binary Matrix $\boldsymbol{\Theta} = \{0,1\}^{\tilde{N}\times K}$ is introduced with elements

$$\Theta_{\tilde{n}k} = \begin{cases} 1 & \text{if a small cell of type } \tilde{n} \text{ is installed in candidate site } k \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

Usually it can be assumed that $\Theta_{\tilde{n}k} = 0 \;\forall\; \tilde{n}, k \in \mathcal{C}^{\mathrm{MC}}$, i.e. no SCs are deployed in the exact cell location of MCs. The cost of deploying small cell type $\tilde{n}$ shall in the following be denoted as $\chi_{\tilde{n}}^{\mathrm{SC}}$. Furthermore, let χ_k^{LOC} denote a cost factor associated with deployment location of cell k, such that the total cost of deploying an SC of type

$\tilde{n}$ in the location of cell k is determined as $\chi_{\tilde{n}}^{\mathrm{SC}}\chi_{k}^{\mathrm{LOC}}$. The total available budget for deployment of SCs shall be $\overline{\chi}$. Further denote as $\varpi_{\tilde{n}}$ the transmit power of SC type $\tilde{n}$.

4.2.1 Greedy Algorithm

An intuitive approach to solve the SC deployment problem is to perform an iterative greedy upgrade approach that starts with the MC-only network and iteratively chooses suitable "upgrades" where either a candidate site is upgraded to a deployed small cell or a deployed small cell is upgraded to a higher-powered model. Assuming that the SC deployment configuration is fully represented by $\boldsymbol{\Theta}$, denote as Θ^{0} and Θ^{UP} the SC deployment solution representing the configuration before and after a considered upgrade. The iterative upgrade process can be summarized as the following:

1. evaluate maximum load of the network before and after upgrade, respectively $\max_k \rho_k\left(\Theta^{0}\right)$ and $\max_k \rho_k\left(\Theta^{\mathrm{UP}}\right)$, based on Eq. (2.10)
2. compute upgrade cost as the difference of total SC configuration costs before and after upgrades: $\sum_{\tilde{n}=1}^{\tilde{N}}\sum_{k\in\mathcal{C}^{\mathrm{SC}}}\chi_{\tilde{n}}^{\mathrm{SC}}\Theta^{\mathrm{UP}}\chi_{k}^{\mathrm{LOC}} - \sum_{\tilde{n}=1}^{\tilde{N}}\sum_{k\in\mathcal{C}^{\mathrm{SC}}}\chi_{\tilde{n}}^{\mathrm{SC}}\Theta^{\mathrm{UP}}\chi_{k}^{\mathrm{LOC}}$
3. the benefit of all possible upgrades is evaluated as the ratio of maximum load decrease to the upgrade cost, and the upgrade with the highest benefit is chosen
4. through the steps 1.-3., the network is iteratively upgraded until the total budget $\overline{\chi}$ is depleted

The proposed greedy upgrade approach maximizes cost efficiency in each upgrade step, but lacks the ability for long-term planning, which is demonstrated and discussed in the simulation results Sec. 4.4.

4.2.2 MILP Formulation

A scheme to solve the small cell deployment problem based on solving a MILP is outlined in the following. The following binary indicator is pre-computed for all possible combinations of cell location k, cell type $\tilde{n}$ and DP m:

$$\Upsilon_{\tilde{n}km} = \begin{cases} 1 \text{ if } \varpi_{\tilde{n}}\Theta_{\tilde{n}k}g_{km} \geq p_j g_j m \;\forall\; j \in \mathcal{C}^{\mathrm{MC}} \\ 0 \text{ otherwise} \end{cases} \tag{4.4}$$

The parameter $\Upsilon_{\tilde{n}km} \in \{0,1\}$ indicated whether a cell of type $\tilde{n}$ deployed in cell location k may provide the strongest signal to DP m. An allocation of DPs to MCs is prevented if an offloading to SCs is possible:

$$\sum_{k \in \mathcal{C}^{\mathrm{MC}}} (1 - A_{mk}) \leq \sum_{j \in \mathcal{C}^{\mathrm{SC}}} \sum_{\tilde{n}}^{\tilde{N}} \Theta_{\tilde{n}k} \Upsilon_{\tilde{n}km} \; \forall \; m \tag{4.5}$$

For the selection of deployed small cell types, represented by $\boldsymbol{\Theta}$, the total budget $\overline{\chi}$ cannot be exceeded:

$$\sum_{\tilde{n}=1}^{\tilde{N}} \sum_{k \in \mathcal{C}^{\mathrm{SC}}} \chi_{\tilde{n}}^{\mathrm{SC}} \chi_{\tilde{n}}^{\mathrm{SC}} \Theta \chi_{k}^{\mathrm{LOC}} \leq \overline{\chi} \tag{4.6}$$

The load balancing problem (2.13) is adapted to account for the possibility of different types of SCs being deployed.

$$\begin{aligned}
\underset{\Pi, \boldsymbol{A}, \boldsymbol{\Theta}}{\text{minimize}} \quad & \Pi && \text{(4.7a)} \\
\text{subject to} \quad & \Pi \geq \sum_{m=1}^{M} A_{km} \frac{d_m}{\eta_{km}^{\mathrm{BW}} W} \zeta_{\tau^{\mathrm{MIN}}}^{+} (\gamma_{km}) \; \forall \; k && \text{(4.7b)} \\
& (4.5), (4.6) \\
& \sum_{k=1}^{K} A_{km} = 1 \; \forall \; m && \text{(4.7c)} \\
& \sum_{\tilde{n}}^{\tilde{N}} \Theta_{\tilde{n}k} \leq 1 \; \forall \; k \in \mathcal{C}^{\mathrm{SC}} && \text{(4.7d)} \\
& \Theta_{\tilde{n}k} = 0 \; \forall \tilde{n}, k \in \mathcal{C}^{\mathrm{MC}} && \text{(4.7e)} \\
& \Pi \in \mathbb{R}_{0+} && \text{(4.7f)} \\
& A_{km}, \Theta_{\tilde{n}k} \in \{0,1\} \; \forall k, m, \tilde{n} && \text{(4.7g)}
\end{aligned}$$

In problem (4.7), the established load balancing problem (2.13) is supplemented with the following constraints to enable the deployment of optimized SC types: equations (4.5) regulates the offloading of DPs to SCs from MCs, (4.6) prevents the deployment solution from exceeding the available budget, and due to (4.7d), only up to one type of SC can be deployed in any given SC candidate site.

The formulation of the deployment problem in (4.7) is linear in all optimization parameters and can be solved using conventional MILP solvers. The feasibility of using this MILP-based approach to obtain near-optimal SC deployment solutions is demonstrated in Sec. 4.4 based on simulated network scenarios.

4.3 Cell Activity Scheduling

For the application of activity scheduling the system model presented in Sec. 2.2 is in the following expanded by a timescale. Using this timescale, interdependent decisions that follow each other, such as scheduling the activity of a cell at the cost of energy consumption, can be adequately modeled. It is assumed that each demand point exhibits a data demand $d_m^{\{t\}}$ in bits per second in time-slot $t = 1, \ldots, T$. Each time-slot t has the time duration l_t. The path attenuation of cell k serving users in pixel m in time-slot t is denoted as $g_{km}^{\{t\}}$, which includes the antenna gains of the base station and user antennas and the propagation loss. Considering only the interference of macro cells, the signal-to-interference-and-noise-ratio (SINR) of the wireless link between cell k and users in pixel m can be formulated as

$$\gamma_{km}^{\{t\}} = \frac{p_k g_{km}^{\{t\}}}{\sum_{j \in \mathcal{C}^{\mathrm{MC}}, j \neq k} p_j g_{jm}^{\{t\}} + \sigma^2}, \tag{4.8}$$

where p_k denotes the transmit power of cell k and σ^2 is the power of additive white Gaussian noise. Note that interference from small cells is omitted because of their much lower transmit power and to simplify the resulting optimization problem. The fraction of its available resources cell k utilizes in order to serve the data demand $d_m^{\{t\}}$ of users in pixel m at time-slot t is characterized as $d_m^{\{t\}}/(\eta_{km}^{\mathrm{BW}} W \log_2(1+\gamma_{km}^{\{t\}}))$ Denote the binary matrix $\boldsymbol{A}^{\{t\}} \in \{0,1\}^{K \times M}$ with its elements defined as $A_{km}^{\{t\}} = 1$ if pixel m is allocated to cell k in time-slot t and $A_{km}^{\{t\}} = 0$ otherwise. The set of allocation matrices for all time-slots is denoted as $\mathcal{A} = \left\{\boldsymbol{A}^{\{1\}}, \ldots, \boldsymbol{A}^{\{T\}}\right\}$. The allocation rule in which each pixel is allocated to the cell providing the strongest signal is given by Eq. (2.14). The total load incurred by cell k at time-slot t, as a ratio of used and available resources in the cell, can be computed as

$$\rho_k^{\{t\}}\left(\boldsymbol{A}^{\{t\}}\right) = \sum_{m=1}^{M} A_{km}^{\{t\}} \frac{d_m^{\{t\}}}{\eta_{km}^{\mathrm{BW}} W \log_2(1+\gamma_{km}^{\{t\}})}. \tag{4.9}$$

The vector of all cell loads at time-slot t is defined as $\boldsymbol{\rho}^{\{t\}} \in \mathbb{R}_{0+}^{K \times 1}$ and the set of all load vectors as $\mathcal{R} = \left\{\boldsymbol{\rho}^{\{1\}}, \ldots, \boldsymbol{\rho}^{\{T\}}\right\}$. To indicate the activity of cell k at time-slot t the binary vector $\boldsymbol{b}^{\{t\}} \in \{0,1\}^{K \times 1}$ is introduced with its k-th element defined as

$$z_k^{\{t\}} = \begin{cases} 1 & \text{if cell } k \text{ is active in time-slot } t \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

The set of all activity indicator vectors is denoted as $\mathcal{Z} = \left\{\boldsymbol{z}^{\{1\}}, \ldots, \boldsymbol{z}^{\{T\}}\right\}$.
Based on the introduced parameters, the amount of stored energy of small cell $k \in \mathcal{C}^{\mathrm{SC}}$ in time-slot t is modeled according to its on-off activity status $z_k^{\{t\}}$ and its load level $\rho_k^{\{t\}}$ as

$$E_k^{\{t\}}(\mathcal{R}, \mathcal{Z}) = E_k^{\{0\}} + \sum_{t'=1}^{t-1} E_k^{\{t'\}} - \sum_{t'=1}^{t} z_k^{\{t'\}} l_{t'} \left(P^{\mathrm{ON}} + P^{\mathrm{LD}} \rho_k^{\{t'\}}\right). \tag{4.11}$$

In Eq. (4.11), $E_k^{\{0\}}$ refers to the energy stored by cell k at the beginning of the observed time horizon, P^{ON} represents a fixed power consumption that is due if the cell is active and P^{LD} is the weighting factor of the energy consumption that scales linearly with the cell load.

4.3.1 Energy and Activity Management

In the following a scheme is introduced to perform load balancing in a heterogeneous wireless network by scheduling the activity of energy harvesting small cells over a time horizon. Without loss of generality it is assumed that energy management and activity scheduling is required only for the small cells in the network. A mixed-integer nonlinear optimization problem (MINLP) can be formulated as follows:

$$\begin{aligned}
\underset{\mathcal{A},\mathcal{Z},\mathcal{R},\Pi}{\text{minimize}} \quad & \Pi && \text{(4.12a)}\\
\text{subject to} \quad & \rho_k^{\{t\}}\left(\boldsymbol{A}^{\{t\}}\right) \leq \Pi \quad \forall k, t && \text{(4.12b)}\\
& E_k^{\{t\}}(\mathcal{R}, \mathcal{Z}) \geq 0 \quad \forall k \in \mathcal{C}^{\mathrm{SC}}, t && \text{(4.12c)}\\
& \sum_{k=1}^{K} A_{km}^{\{t\}} \delta_k p_k g_{km}^{\{t\}} \geq \\
& \qquad \left(1 - A_{jm}^{\{t\}}\right) \delta_j p_j g_{jm}^{\{t\}} \quad \forall j, m, t && \text{(4.12d)}\\
& \sum_{k=1}^{K} A_{km}^{\{t\}} = 1 \quad \forall m, t && \text{(4.12e)}\\
& A_{km}^{\{t\}} \leq z_k^{\{t\}} \quad \forall k \in \mathcal{C}^{\mathrm{SC}}, m, t && \text{(4.12f)}\\
& \Pi \in \mathbb{R}_{0+},\ \Pi \leq \overline{\Pi} && \text{(4.12g)}\\
& \rho_k^{\{t\}} \in \mathbb{R}_{0+} \quad \forall k, t && \text{(4.12h)}\\
& A_{km}^{\{t\}}, z_k^{\{t\}} \in \{0, 1\} \quad \forall k, m, t && \text{(4.12i)}
\end{aligned}$$

In problem (4.12), the objective (4.12a) and constraints (4.12b) minimize the maximum load Π occurring for any cell in any time-slot, which is commonly referred to as load

balancing. This maximum load level is upper bounded in (4.12g) by $\Pi \leq \overline{\Pi}$, which is typically set to $\overline{\Pi} = 1$. Load levels greater than one indicate overloaded cells, making the network configuration technically infeasible. The constraints (4.12c) ensure that each small cell cannot utilize more than its available energy, i.e. the energy level at the end of each time-slot is nonnegative. Constraints (4.12d) are a reformulation of (2.14), and due to (4.12e) and (4.12f), each pixel is allocated to an active cell.
Problem (4.12) is a nonlinear mixed-integer programming problem because of the bilinear products $z_k^{\{t'\}}\rho_k^{\{t\}}$ in the constraints (4.12c) where $E_k^{\{t\}}(\mathcal{R}, \mathcal{Z})$ is computed as given in Eq. (4.11). Since the load levels are bounded in problem (4.12) by $\rho_k^{\{t\}} \leq \Pi \leq \overline{\Pi}$, the products of binary and a continuous parameters can be reformulated by applying a lifting strategy and introducing the auxiliary parameter $\tilde{\rho}_k^{\{t\}}$ with the following set of inequalities:

$$\tilde{\rho}_k^{\{t\}} \leq \rho_k^{\{t\}} \quad \forall k, t \tag{4.13a}$$
$$\tilde{\rho}_k^{\{t\}} \leq z_k^{\{t'\}}\overline{\Pi} \quad \forall k, t \tag{4.13b}$$
$$\tilde{\rho}_k^{\{t\}} \geq \rho_k^{\{t\}} - \left(1 - z_k^{\{t'\}}\right)\overline{\Pi} \quad \forall k, t. \tag{4.13c}$$

Further denote as $\tilde{\boldsymbol{\rho}}^{\{t\}} \in \mathbb{R}_{0+}^{K\times 1}$ the vector of elements $\tilde{\rho}_k^{\{t\}}$ for all k, and the set $\mathcal{Y} = \left\{\tilde{\boldsymbol{\rho}}^{\{1\}}, \ldots, \tilde{\boldsymbol{\rho}}^{\{T\}}\right\}$. Replacing (4.12c) by (4.13), the problem (4.12) can therefore be reformulated into a the following mixed-integer linear problem (MILP):

$$\underset{\mathcal{A},\mathcal{Z},\mathcal{R},\mathcal{Y},\Pi}{\text{minimize}} \quad \Pi \tag{4.14a}$$
$$\text{subject to} \quad (4.12\text{b}), (4.13), (4.12\text{d}) - (4.12\text{g})$$
$$0 \leq E_k^{\{0\}} + \sum_{t'=1}^{t-1} E_k^{\{t'\}} - \sum_{t'=1}^{t}\left(z_k^{\{t'\}} l_{t'} P^{\text{ON}} + \tilde{\rho}_k^{\{t'\}} l_{t'} P^{\text{LD}}\right) \forall k, t \tag{4.14b}$$
$$\rho_k^{\{t\}}, \tilde{\rho}_k^{\{t'\}} \in \mathbb{R}_{0+} \quad \forall k, t \tag{4.14c}$$
$$A_{km}^{\{t\}}, z_k^{\{t\}} \in \{0, 1\} \quad \forall k, m, t \tag{4.14d}$$

The SC activity scheduling problem as defined in Eq. (4.14) is a MILP, for which computationally efficient state-of-the-art solvers are available as discussed in Sec. 3.2.1. Even though this is a computationally tractable approach for the SC activity scheduling problem, the increased dimensionality obtained from the lifting procedure may prove problematic if a high number of time-slots are jointly optimized. The following Sec. 4.3.2 addresses this challenge by decreasing the problem size along the timescale dimension.

4.3.2 Timescale Optimization

The approach proposed in Sec. 4.3.1 relies on solving an optimization problem that can be computationally challenging if a large number of time-slots T are jointly optimized. In the following a scalable approach is proposed in which multiple forecast snapshots are grouped together into a smaller number of time-slots. The optimization procedure proposed in Sec. 4.3.1 is applied to the resulting lower number of time-slots. Suppose that there exist S network snapshots indicated by $s = 1, \ldots, S$ which represent a demand forecast that is valid for time duration $\tilde{l}_s$. In snapshot s, pixel m is forecasted to exhibit an average aggregated data demand $\tilde{d}_m^{\{s\}}$. Let $\boldsymbol{J} \in \{0,1\}^{S \times T}$ be a matrix indicating the grouping of snapshots s to time-slots t with its elements indicated as $J_{s,t} = 1$ if time-slot t contains snapshot s and $J_{s,t} = 0$ otherwise. Therefore there is $l_t = \sum_{s=1}^{S} \tilde{l}_s J_{s,t}$ and

$$d_m^{\{t\}} = \frac{1}{l_t} \sum_{s=1}^{S} \tilde{l}_s \tilde{d}_m^{\{s\}} J_{s,t}. \tag{4.15}$$

The mean squared deviation of the user demand in all pixels between snapshots s and $s-1$ can be computed as

$$v(s) = \frac{\alpha}{M} \sum_{m=1}^{M} \left(\tilde{d}_m^{\{s\}} - \tilde{d}_m^{\{s-1\}} \right)^2 \quad \forall s > 1 \tag{4.16}$$

where α is a scaling parameter that is chosen such that $\max_s v(s) = 1$. The function $v(s)$ in Eq. (4.16) measures the average squared difference in pixel demand between $s-1$ and s. The proposed strategy to find an optimized allocation matrix $\boldsymbol{J}$ of snapshots to time-slots is to use a high density of snapshots when $v(s)$ takes high values, indicating a high demand fluctuation in the network, and a low density of snapshots when $v(s)$ takes lower values, indicating that the pixel demands remain mostly unchanged. This is achieved by solving the following optimization problem:

$$\underset{\Psi, \boldsymbol{J}}{\text{minimize}} \quad \Psi \tag{4.17a}$$

$$\text{subject to} \quad \sum_{s=1}^{S} J_{s,t} v(s) \leq \Psi \quad \forall t \tag{4.17b}$$

$$\sum_{s=1}^{S} J_{s,t} \geq 1 \quad \forall t \tag{4.17c}$$

$$\sum_{t=1}^{T} J_{s,t} = 1 \quad \forall s \tag{4.17d}$$

$$J_{1,1} = 1, J_{S,T} = 1 \tag{4.17e}$$

$$J_{s,t} \leq J_{s-1,t} + J_{s-1,t-1} \; \forall s > 1, t \tag{4.17f}$$

$$\Psi \in \mathbb{R}_{0+}, J_{s,t} \in \{0,1\} \quad \forall s, t, S > T \tag{4.17g}$$

In problem (4.17), the constraints (4.17b) use the term $v(s)$ as a cost function, where the maximum sum cost allocated to any time-slot is minimized. With constraints (4.17c) and (4.17d), at least one snapshot is allocated to each time-slot and each snapshot can only be allocated to one time-slot, respectively. Constraints (4.17e) cause the first and last snapshot to be allocated to the first and last time-slot respectively, and due to (4.17f), starting from time-slot $t = 1$, with increasing s, each snapshot can only be allocated to the current time-slot (t) or the next time-slot ($t + 1$).

4.4 Simulation Results

A simulation for a heterogeneous wireless communication network is carried out using the parameters listed in Table 4.4 and the locations of 3 MCs and 9 SCs as depicted in Fig.4.4. The network area is segmented into pixels of size 25×25 m and the number of users U_m in pixel m is determined randomly in each snapshot from a Poisson distribution $U_m \sim \mathfrak{P}(\lambda_m)$ with rate $\lambda_m = 0.1$ for normal pixels and higher λ_m if pixel m is in a hotspot. This results in about 300 users in the simulated network area. The used performance criterion is the maximum load of any cell during any time-slot, which has to be minimized. Interferences and load levels are computed considering all macro cells and active small cells, which is an exact computation of the interference contrary to the simplifying approximation of neglecting the SC interference that was used in the optimization schemes discussed in Secs. 4.2 and 4.3.

Two simulations are performed to validate the performance of the methods introduced in Secs. 4.2.1 and 4.2.2. Common parameters for both simulations are shown in Table 4.4. The costs of small cell models, as well as the total budget, are modeled in 'Units'. The number of users and their data demand is chosen in such a way that the macro cells are close to being overloaded, with a load factor $0.8 < \rho_k < 1$. The MILP problems are solved using Gurobi Optimizer 6.0 [GUR] and the CVX optimization toolbox for MATLAB [GB14]. As the main performance metric, the 'relative maximum load' is computed, which is the maximum load factor of any cell in the network after small cell deployment normalized by the maximum load factor in the original macro-only network. Each curve in Figs. 4.2 and 4.3 is obtained from averaging 20 Monte-Carlo trials, each generated from a new map with random hotspot and DP distributions.

Table 4.1. Common network parameters for the simulation of a heterogeneous LTE network.

Area size	1000×1000 m
System bandwidth W	20 MHz
Avg. number of users	≈ 350
Avg. demand per user	400 kbit/s
Log-normal shadow fading	5 dB
Noise power	-145 dBm/Hz
Bandwidth efficiency η^{BW}	0.8
MC transmit power p_k	46 dBm
MC antenna gain $\tilde{g}^{\mathrm{ABS}}$	15dB
Propagation loss g^{MC}, g^{SC}	3GPP TS 36.814 [3GP16]

Simulation 1 is designed to evaluate the two proposed deployment methods, MILP-based and greedy approach, in comparison with the optimal solution found with exhaustive search. Finding this optimal solution is only possible for a very small instance of the problem. The simulation parameters are shown in Table 4.4. Only ten candidate sites are used, and the two solutions for each candidate site are either 'no deployment' or deployment of a pico cell with 1 W transmit power and 5 dB antenna gain, which will in the following be called 'pico B'. The decrease in the maximum load factor is evaluated for multiple small total budgets up to $Z = 300$.

As observable in Fig. 4.2, the decrease in load factor is less than 15% for all methods, due to the limited budget and small pico cell model used. The relative maximum load decreases continuously with the increasing budget for the exhaustive search approach. Since the greedy approach is iterative in nature and the MILP-based approach only solves an approximated version of the original problem, both methods show slightly worse performance than the optimal results, but provide good solutions for the small cell deployment problem. Simulation 2 compares the two proposed methods for a more realistic size of the deployment problem. As shown in Table 4.4, the number of candidate sites is increased to 100, and 4 different small cell deployment options are available with different cost and transmit power. An example for the resulting deployment solution is illustrated in Fig. 4.1. The performance comparison of the CWGU and the MILP-based approach for Simulation 2 is shown in Fig. 4.3. The decrease in maximum load is much larger than the decrease obtained in Simulation 1, because of the larger budget and the increased number of small cell models. The MILP-based approach achieves a larger decrease in terms of maximum load than the CWGU. Since the computation time for the MILP is also much lower with the used system, this approach is preferable over the iterative CWGU method.

Table 4.2. Hotspot model, deployment cost factors and small cell models for SC deployment simulation.

Pixel size	25×25 m
Number of hotspots	10
Hotspot radius	50m
Hotspot user density	4× normal
Area deployment cost factors	1 in 50% of area 0.75 in 15% of area 1.5 in 25% of area 3 in 10% of area
Pos. of macro BS	MC1 at [100m, 200m] MC2 at [800m, 100m] MC3 at [200m, 800m] MC4 at [900m, 900m]
number of candidate sites (simulation 1)	10
small cell models (simulation 1):	
no deployment	$\varpi_1 = 0$, $\chi_1^{\mathrm{SC}} = 0$ Units
pico C	$\varpi_2 = 35$dBm, $\chi_2^{\mathrm{SC}} = 50$ Units
number of candidate sites (simulation 2)	100
small cell models (simulation 2):	
no deployment	$\varpi_1 = 0$, $\chi_1^{\mathrm{SC}} = 0$ Units
pico C	$\varpi_2 = 29$dBm, $\chi_2^{\mathrm{SC}} = 40$ Units
pico B	$\varpi_3 = 35$dBm, $\chi_3^{\mathrm{SC}} = 50$ Units
pico A	$\varpi_4 = 41$dBm, $\chi_4^{\mathrm{SC}} = 75$ Units

Table 4.3. Small cell energy management and activity scheduling simulation parameters

Bias values SC $\delta_k \forall k \in \mathcal{C}^{\mathrm{SC}}$	6dB
Initial energy $E_k^{\{0\}} \forall k \in \mathcal{C}^{\mathrm{SC}}$	2
Fixed power cons. fac. P^{ON}	1
Load-dependent power cons. fac. P^{LD}	1
Snapshot length $\tilde{l}_s \forall s$	1
Number of snapshots S	48

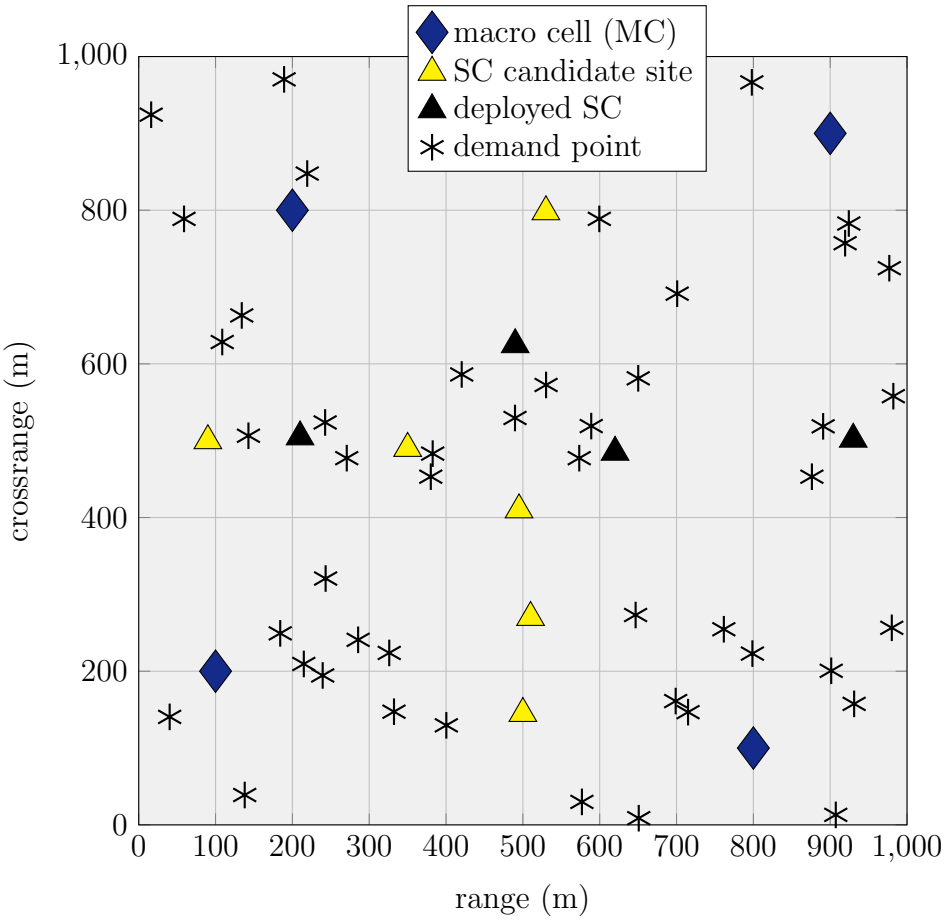

Figure 4.1. Network scenario and sample solution for SC deployment panning simulation. Different types of SC are deployed on the edges between the MC coverage areas.

The methods used for benchmarking the proposed small cell activity scheduling schemes are the following: The first method is to leave all SCs off and handle data traffic solely through macro cells. Since this solution is part of the feasible set of problem (4.14), and the only simplification of neglecting small cell interferences is not very significant, leaving all SCs off always generates higher load levels than the proposed scheme. The second benchmarking method is to leave all small cells on at all times, which ignores the constraints (4.14b), but serves as an upper bound on achievable performance. The third benchmarking method is to ignore (4.14b) and find the best activity schedule for each snapshot independently using exhaustive search. This method may serve as the absolute theoretical lower bound on achievable load levels that is however computationally impractical for larger networks. For each of the following three simulations, 100 network scenarios with $S = 48$ demand forecast snapshots are simulated, and the resulting maximum cell loads are averaged for all methods.

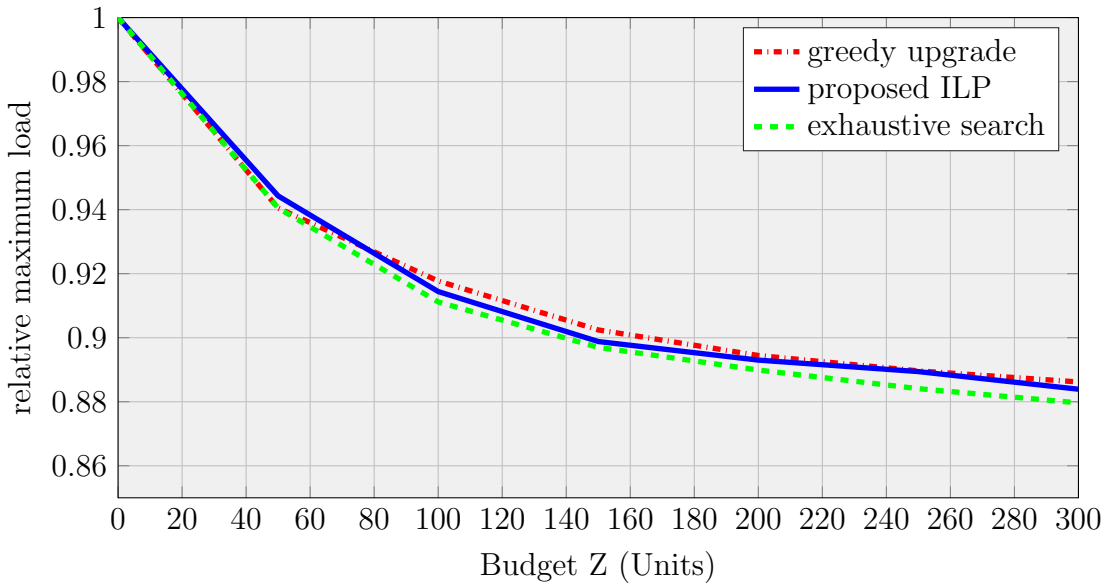

Figure 4.2. Small cell deployment performance (simulation 1), with low number of candidate sites and only one small cell type, to allow for comparison with the exhaustive search solution.

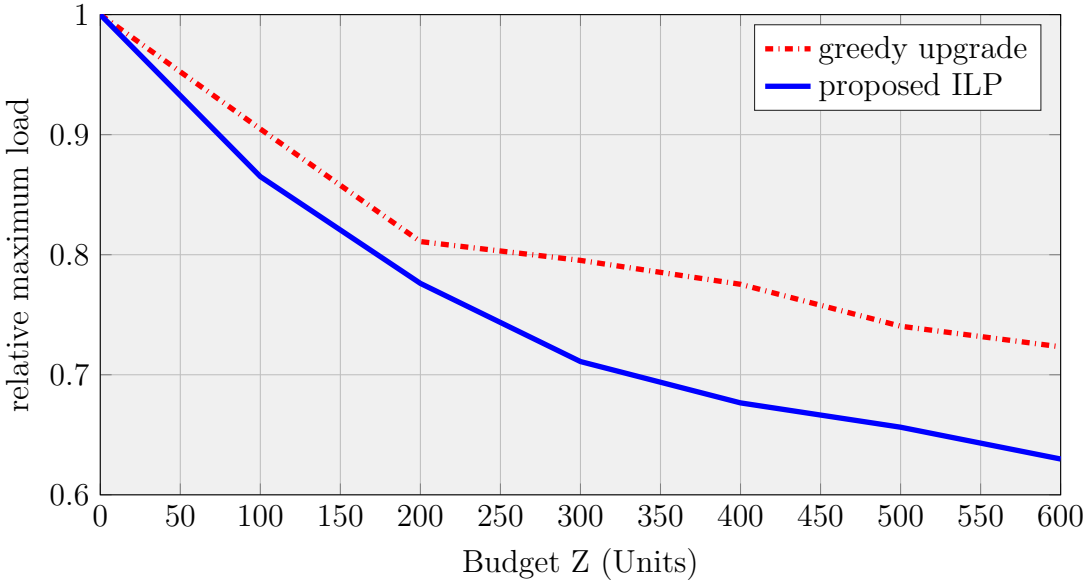

Figure 4.3. Small cell deployment performance (simulation 2) with a large number of candidate sites and three selectable small cell types.

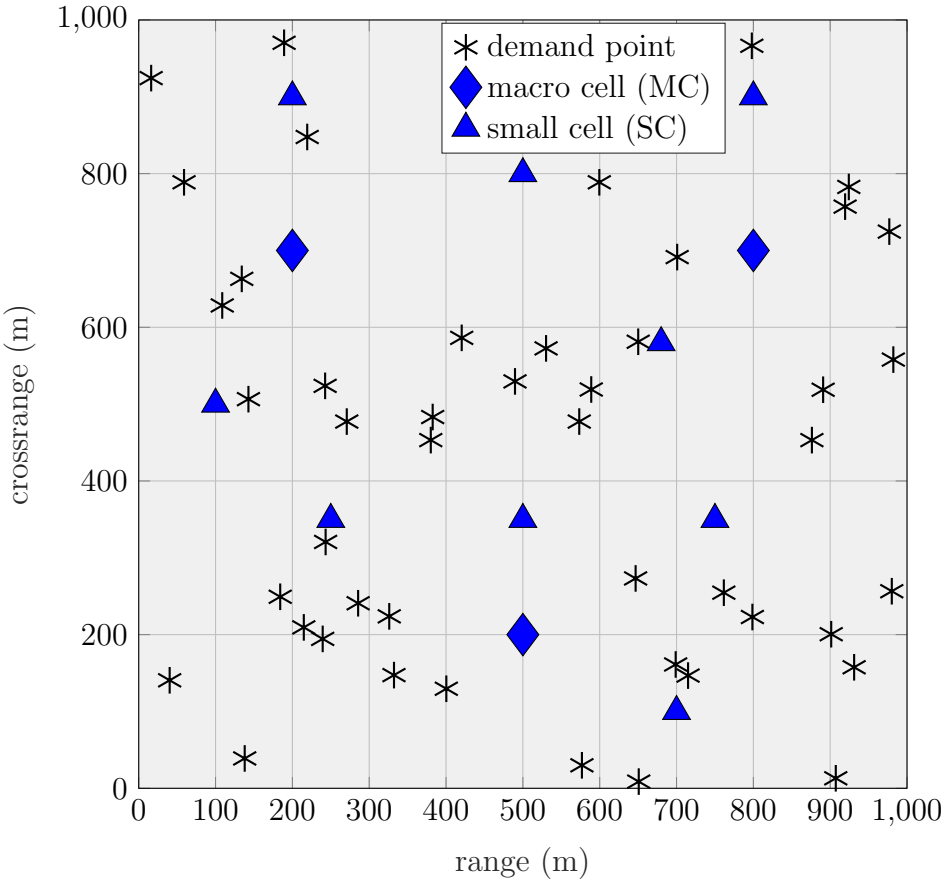

Figure 4.4. Network scenario for SC scheduling simulation. Three MC and nine SC are deployed in the network area under consideration, all SC are assumed to utilize a renewable energy source and energy storage.

The results of a simulation with variable DP demand are depicted in Fig. 4.5, where the maximum load for all methods is shown over increasing user data demand and with varying incoming energy levels $E_k^{\{t\}}$ for the proposed method. Every 8 snapshots, the coverage areas of three randomly selected small cells are chosen as hotspots. Problems (4.14) and (4.17) are solved with $T = 8$ time-slots. As the amount of harvested energy increases, more small cells can be left on in the proposed approach, and the load level decreases. In a simulation of the achieved load level over a variable number of timeslots, shown in Fig. 4.6, the demand of a single user is fixed at 400 kbit/s. The maximum load level over the number of time-slots T is shown in Fig. 4.6. It is observable that the achieved maximum load decreases with increasing $E_k^{\{t\}}$ as well as with an increasing number of time-slots T.

To highlight the benefits of the proposed timescale optimization technique proposed with problem (4.17), a network scenario is constructed that has a high demand fluc-

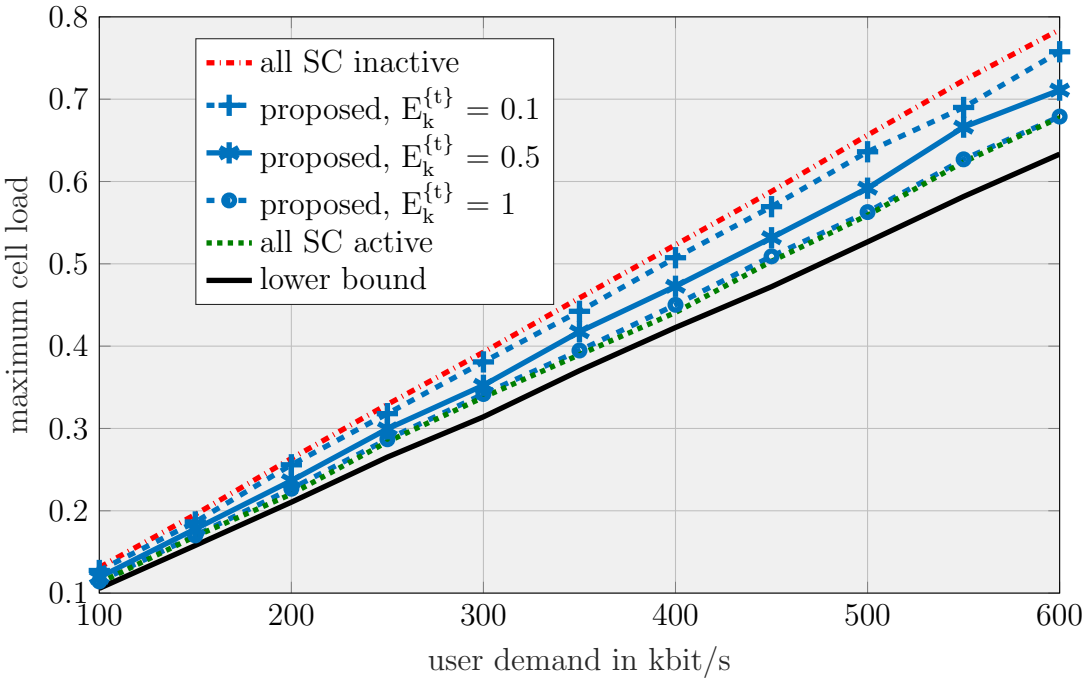

Figure 4.5. Averaged maximum load level for different small cell scheduling approaches and varying amounts of energy supply for the SC. The achieved load levels increase linearly with the user demand.

tuation. For the previous simulations, the overall demand fluctuations, represented by the function $v(s)$ in Eq. (4.16), is chosen to be very low, leading to an almost equal length of all time-slots. For a network scenario with high demand variability, a network scenario is constructed where, over the $S = 48$ snapshots considered, three new hotspots are selected randomly every 4 snapshots for $s = [15, 35]$. Additionally, the overall data demand is multiplied by a factor of 1.5 for $s = [15, 21]$ and $s = [29, 35]$ and by a factor of 2 for $s = [22, 28]$. This leads to an increased cost function $v(s)$ for $s = [15, 35]$ shown in Fig. 4.7.

The resulting segmentation of snapshots into $T = 8$ time-slots obtained from solving problem (4.17) is also shown in Fig. 4.7. As observable, a higher density of time-slots with shorter duration each is chosen in the high variability time interval. The comparison in this network scenario between the proposed approach and the approach where a uniform length of time-slots is chosen is shown in Fig. 4.8. The propsed timescale optimization achieves significantly lower load levels especially if a low number of time-slots is available. The proposed joint optimization of the small cell activity schedule and of the time-slot durations on which the schedule is applied achieves significantly decreased load levels.

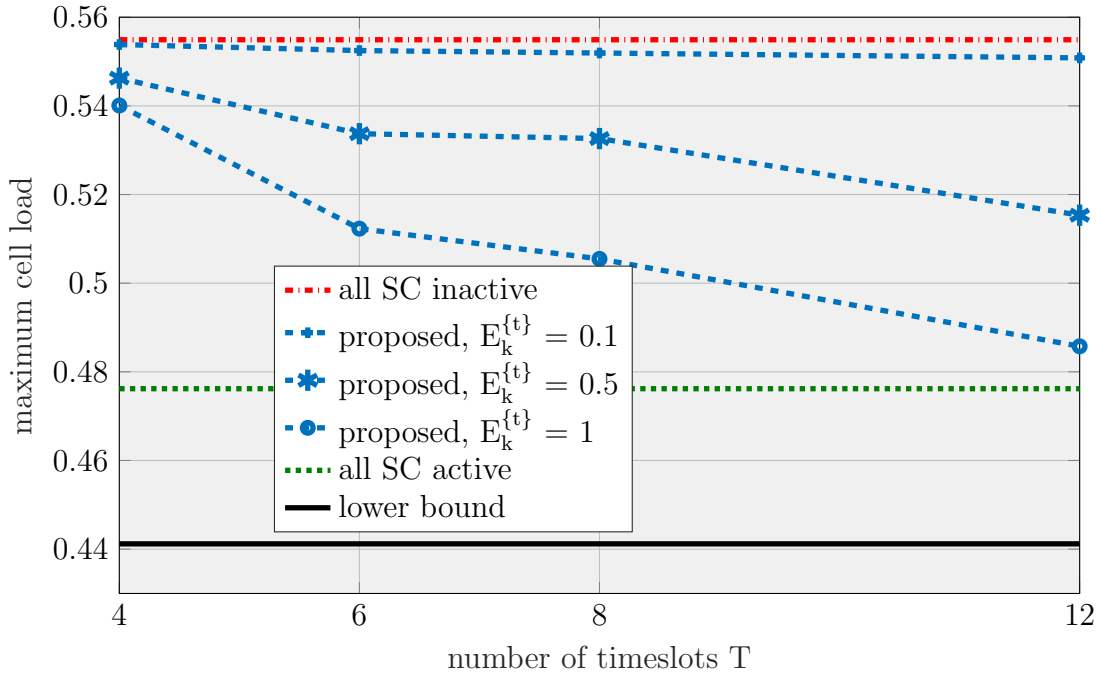

Figure 4.6. Averaged maximum load levels for different number of time-slots and varying energy supply. The maximum load level decreases if more time-slots can be jointly optimized.

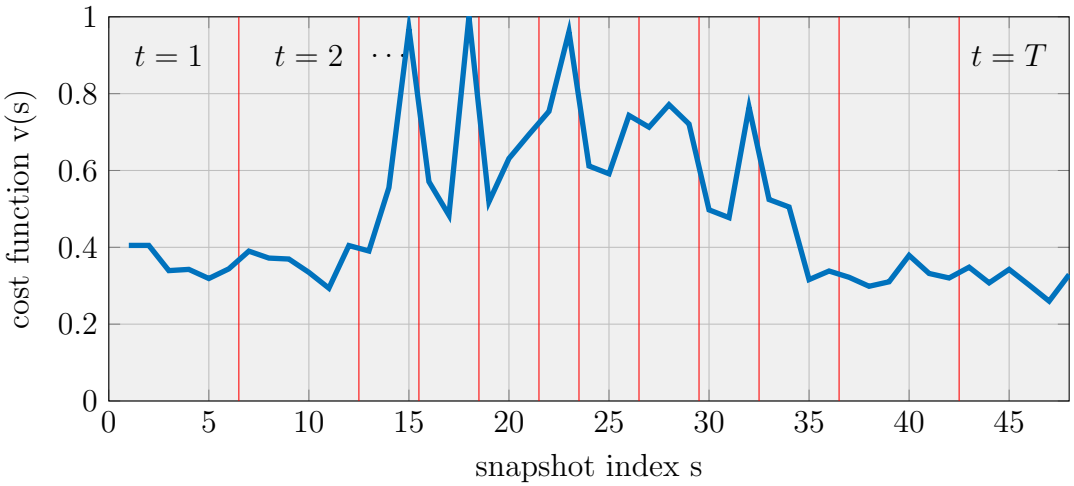

Figure 4.7. Snapshot cost function example with corresponding time-slot segmentation. A time period with high demand variability in the network was added for snapshots 15 to 35.

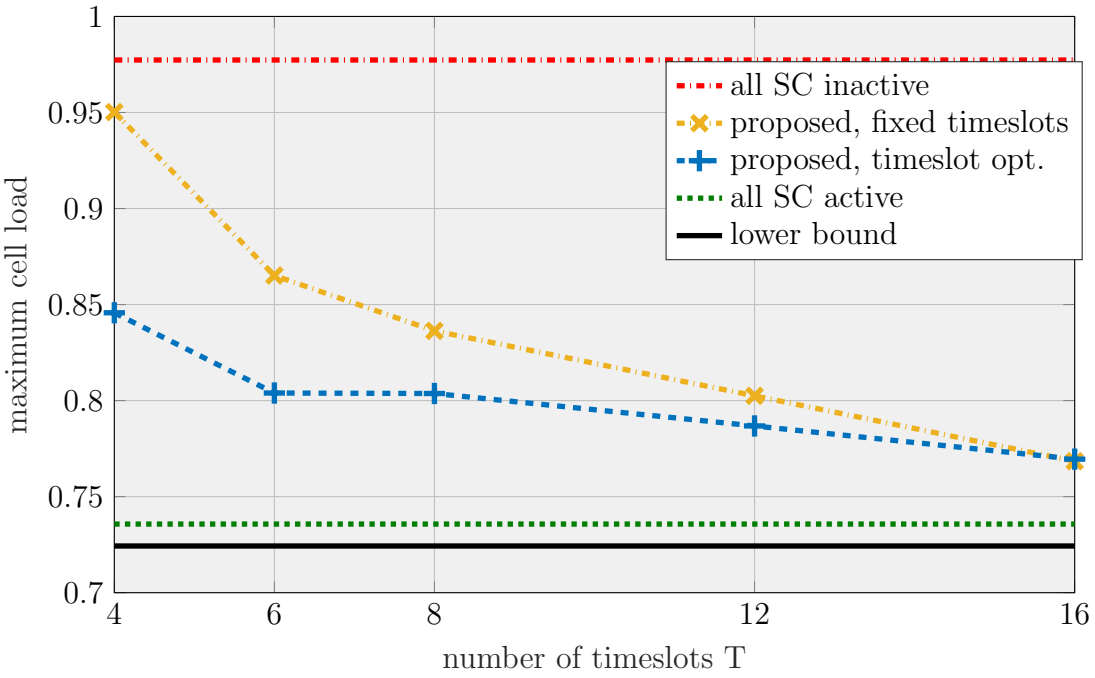

Figure 4.8. Averaged maximum load over number of time-slots, with fixed and varying time-slot length. The proposed approach with variable time-slot length achieves lower load levels than with fixed time-slot duration, especially if a low number of time-slots can be jointly optimized.

4.5 Summary

A framework was introduced for the planning of SC deployments and the scheduling of their activity status over a time horizon. The proposed SC deployment scheme considers multiple heterogeneities such as hotspots in the mobile user distribution, location-dependent site acquisition costs and multiple available small cell models with different cost and transmit power. An approach to find candidate deployment locations with a 'site suitability function' has been introduced. Due to the non-linearity of the original mixed-integer optimization problem, a greedy upgrade method to solve the problem iteratively, and a MILP-based approximation of the original problem, which can be solved using conventional optimization tools. The superiority of the MILP-based approach was demonstrated with simulations of a heterogeneous LTE network. A scheme for small cell activity scheduling over a given time horizon with energy harvesting small cells was introduced. The optimal activity schedule was obtained as the

result of a mixed-integer optimization problem. To achieve scalability of the mixed-integer optimization approach, a time schedule with time-slots of varying length, on which the schedule was applied, was optimized according to the temporal demand variability of the network. Simulation results show that the proposed approach achieves a decrease in load levels and problem dimensionality. The deployment, activity scheduling, and timeslot optimization subproblems were all effectively solved to facilitate a load balanced network already at the network planning stage.

Chapter 5

Resource Allocation and Network Slicing

5.1 Introduction

As a key feature, future wireless networks are designed to operate based on a "Network Slice Layer" and a "Service Layer" [P1 16,FMK17,FSPRA18,SPRFA17]. The function of a network slice is to aggregate sets of network resources from the underlying physical layer to provide specific services in the service layer [FSPRA18]. These services, which can be categorized for example into eMBB, URLLC and mMTC services as discussed in Sec. 1.1, can have varying QoS-constraints such as a minimum SINR level, bandwidth efficiency, latency, support for a large number of connected devices and other requirements. The QoS-constraints are largely dictated by the requested service, and the underlying time-frequency resources have to be distributed accordingly. Methods for spectrum coexistence between cells aim to find solutions where each cell utilizes as much of the spectrum as required, but adapts its spectrum-sharing behavior in a "cognitive" manner based on that of neighboring cells, to avoid interfering in critical scenarios [ADARCA17]. The segmentation and allocation of the network resources can be planned in advance, based on demand forecasts, or while the network is in operation. As a consequence, there exists multiple approaches to RAN slicing based on spectrum planning, inter-cell interference coordination, packet scheduling or admission control [SPRFA17]. The approach proposed this chapter is a spectrum planning scheme, but its solutions can be applied to admission control schemes.

From a network operators perspective one of the greatest concerns is how to enable the coexistence of a variety of services, with very diverse requirements regarding reliability, data throughput or latency, in a network that due to its dense cell deployment becomes increasingly interference-limited [AZDG16]. To achieve the high SINR levels that are necessary for high-reliability or high-throughput slices, the network can only be "densified" to a limited extent if co-channel operation is being considered, as discussed in Sec. 1.1. The resource allocation scheme proposed in this chapter aims to minimize the amount of resources that have to be utilized to fulfill all QoS-requirements of a given network scenario. This is achieved by a joint optimization of the resource distribution between slices and the allocations these resources to cells and DPs.

5.1.1 State-of-the-Art

The approach of segmenting the time-frequency resources and allocating these segments to cells originates from conventional cell planning [EHF08], but can be adapted to optimize resource allocation to slices providing different services. To mitigate the significant interferences caused by network densification, multiple slices operating on orthogonal sets of physical time-frequency resources have been proposed [ZLC+17, P1 16]. The authors in [ADARCA17] propose a spectrum managements scheme where the spatial distribution of cells is considered, and resources are distributed according to service requirements. There is however no method provided to optimize the resource efficiency or resource consumption. In [ZLC+17], a scheme is introduced for handover management and for maximizing individual SC capacities, but the proposed methods a not applicable for spectrum management of a larger network and joint optimization of multiple cells. Dedicated network optimization methods for dense heterogeneous slicing networks are still scarce in current literature [AZDG16].

5.1.2 Contributions and Overview

In this work, an approach is proposed for network time-frequency resource planning based on maximizing the resource efficiency of the network by joint optimization of the resource assignment to network slices, the allocation of slices to different operating cells, and the allocation of users to cells. The proposed resource optimization is a concrete application of the concepts of spectrum planning and densification of HetNets formulated in [AZDG16, SPRFA17]. It is also demonstrated how SINR-requirements and bandwidth efficiencies of the transmission schemes dedicated to different services can be incorporated into the network optimization process. The contributions of the proposed resource allocation method compared to state-of-the-art approaches can be summarized as follows: The proposed approach jointly considers multiple network parameters for optimization: the allocation of DPs to cells, the allocation of cells to different time-frequency resource pools and the dimensioning of these resource pools. Established methods consider a static model where the interference for each DP is considered to be fixed. The proposed approach dynamically models the interference based on the jointly optimized resource pools. QOS-constraints for different services provided by the network slices can be considered during the optimization of the resource distribution.

In the following Sec. 5.2, a problem formulation is introduced for maximizing the resource efficiency of a wireless network. The proposed resource allocation method in

is discussed in Sec. 5.3. Simulation results of the achieved resource efficiency for varying user demand and number of small cells are given in Sec. 5.4. A final assessment of the method and summary of the results is provided in Sec. 5.5.

5.2 Problem Formulation

The system model for network slicing expands upon the model introduced in Sec. 2.2. Bandwidth resources available to the network are divided among I slices, indicated by $q = 1, \ldots, Q$. These slices may be designed to provide different services, with varying minimum SINR requirements γ_q^{MIN} and bandwidth efficiencies η_q^{Q}, which represent the ratio of the bandwidth available for data transmission to the total available bandwidth. Other service requirements may include peak data rates and latency, which both can be optimized in higher network layers. A low-latency transmission scheme for example would rely on very small packet sizes, which can be modeled with a decreased bandwidth efficiency in the proposed scheme [MK10]. The information in which of the slices q a DP m can be served is specified by parameter S_{qm}, where $S_{qm} = 1$ if DP m can be served in slice q, and $S_{qm} = 0$ otherwise. In this chapter, it is assumed without loss of generality that each cell operates in a single slice. A cell that operates in multiple slices can be modeled as multiple "virtual" cells in the same location. The allocation of cells to slices is indicated in matrix $\boldsymbol{B} \in \{0,1\}^{Q \times K}$, where

$$B_{qk} = \begin{cases} 1 & \text{if cell } k \text{ operates in slice } q \\ 0 & \text{otherwise} \end{cases} . \tag{5.1}$$

The SINR γ_{km} of DP m served by the base station in cell k is formulated as

$$\gamma_{km} = \frac{p_k^{\mathrm{S}} g_{km}}{\sum_{j \in \{\mathcal{C} \setminus \{k\}\}} \sum_{q=1}^{Q} B_{qk} B_{qj} p_j^{\mathrm{S}} g_{jm} + \lambda} \tag{5.2}$$

where p_k^{S} is the power spectral density of the transmitted signal of cell k, g_{km} is the combined attenuation factor from cell k to user m resulting from antenna gains and path loss and λ is the power spectral density of additive white Gaussian noise. The sum over $j \in \{\mathcal{C} \setminus \{k\}\}$ refers to the set of all cells $j = 1, \ldots, K$ except for $j = k$. The model in (5.2) corresponds to an OFDMA system with full frequency reuse between cells [MK10]. Note that the term $j \in \{\mathcal{C} \setminus \{k\}\} \sum_{q=1}^{I} B_{qk} B_{qj} p_j g_{jm}$ indicates the interference from those cells j which are serving, and therefore interfering, in the same slice q as cell k. Indicate the allocation of DPs to cells with the binary matrix $\boldsymbol{A} \in \{0,1\}^{K \times M}$, with its elements $A_{km} = 1$ if DP m is allocated to cell k, and $A_{km} = 0$ otherwise. The

bandwidth efficiency modifying factor related to the type of cell k (e.g. MC or SC) used for the transmission is indicated as η_k^{K}. Based on the cell load computation outlined in [MK10], cell k is not overloaded if the following condition is satisfied:

$$\sum_{m=1}^{M} A_{km} D_m \zeta_{\tau^{\mathrm{MIN}}}^{+}(\gamma_{km}) \leq \eta_k^{\mathrm{K}} \sum_{q=1}^{Q} \left(B_{qk} \eta_q^{\mathrm{Q}} w_q \right) \ \forall\, k, \tag{5.3}$$

where $\zeta_{\tau^{\mathrm{MIN}}}^{+}(\gamma)$ is defined in Eq. (2.9), w_q is the bandwidth allocated to slice q. The total system bandwidth is denoted as $\overline{w}$, such that for any viable network configuration $\sum_{q=1}^{Q} w_q \leq \overline{w}$.

To maximize the overall resource efficiency of the network, a cell- and slicing- configuration need to be found where the requests from all DP are fulfilled with a minimum amount of bandwidth resources used. This problem is equivalent to maximizing the amount of "unused" resources $Z = \overline{w} - \sum_{q=1}^{Q} w_q$. The practical use of this approach is that the unused spectral resources, after optimizing the network using our proposed approach, can be utilized for example to further improve the data rates of selected users. For this purpose a mixed-integer nonlinear optimization problem (MINLP) can be formulated as follows:

$$\underset{Z,\boldsymbol{w},\boldsymbol{A},\boldsymbol{B}}{\text{maximize}} \quad Z \tag{5.4a}$$

$$\text{subject to} \quad \sum_{q=1}^{Q} w_q + Z = \overline{w} \tag{5.4b}$$

$$\sum_{k=1}^{K} A_{km} = 1 \ \forall\, m \tag{5.4c}$$

$$\sum_{q=1}^{Q} B_{qk} \leq \min\left\{ 1, \sum_{m=1}^{M} A_{km} \right\} \ \forall\, k \tag{5.4d}$$

$$\sum_{k=1}^{K} A_{km} B_{qk} \leq S_{qm} \ \forall\, q, m \tag{5.4e}$$

$$\gamma_{km} = \frac{p_k^{\mathrm{S}} g_{km}}{\sum_{j \in \{\mathcal{C} \setminus \{k\}\}} \sum_{q=1}^{Q} B_{qk} B_{qj} p_j^{\mathrm{S}} g_{jm} + \lambda} \ \forall\, k, m \tag{5.4f}$$

$$A_{km}\gamma_{km} \geq A_{km}\gamma_q^{\mathrm{MIN}} \ \forall\, k, \left\{ m, q \middle| \sum_{k} A_{km} B_{qk} = 1 \right\} \tag{5.4g}$$

$$\sum_{m=1}^{M} A_{km} D_m \zeta_{\tau^{\mathrm{MIN}}}^{+}(\gamma_{km}) \leq \eta_k^{\mathrm{K}} \sum_{q=1}^{Q} \left(B_{qk} \eta_q^{\mathrm{Q}} w_q \right) \ \forall\, k \tag{5.4h}$$

$$Z, w_q \in \mathbb{R}_{0+} \ \forall\, q \tag{5.4i}$$

$$A_{km}, B_{qk} \in \{0,1\} \; \forall \; q,k,m \tag{5.4j}$$

In problem (5.4), equality (5.4c) forces each DP to be allocated to exactly one cell, while inequality (5.4e) allows it to be allocated only to its requested slice(s), as specified by S_{qm}. Constraints (5.4d) cause each cell to operate in at most one slice, and only if it has users allocated to it. With (5.4g), an allocated DP-cell connection needs to fulfill the SINR requirement of the slice requested by the DP.
Problem (5.4) is a nonconvex mixed-integer program. Especially the dependency of the interference-plus-noise term in Eq. (5.4f) on $\boldsymbol{B}$ and multiple bilinear terms in other constraints render the problem computationally intractable to solve. Motivated by the considerations in Sec. 3.2.1, an approach to reformulate problem (5.4) into a MILP is derived in the following.

5.3 Resource Planning Scheme

This inner approximation is performed in three steps: firstly, the interference for each connection is upper bounded with a set of discrete interference scenarios, secondly the SINR- and load-computation are reformulated with said approximation, and finally all bilinear products of optimization parameters are replaced with equivalent linear formulations using a lifting procedure.
To provide an approximation of the SINR expression as defined in Eq. (5.4f), a set of discrete interference levels Ψ_{nkm} is introduced, indicated by $n = 1, \ldots, N$, which are precomputed for each pair of (k,m), such that they represent the most relevant low- to medium-SINR scenarios. These discrete interference levels are utilized to approximate the denominator of the SINR-term in Eq. (5.4f), in a scheme based on Sec. 3.2.4. To ensure the feasibility of the approximate problem, the full interference scenario, i.e. $\Psi_{1km} = \sum_{j \in \{\mathcal{C} \setminus \{k\}\}} p_j g_{jm} + \epsilon$ is considered as one of the relevant interference scenarios. The second relevant interference level is the scenario where the strongest interfering cell is inactive or operating in another slice, for example $\Psi_{2km} = \Psi_{1km} - \max_{j \setminus \{k\}} p_j g_{jm}$, . Usually the removal of the first- and second-strongest interfering cell has the highest impact on achievable rates. If every possible interference scenario is considered for K cells, 2^K scenarios are required to solve the resource allocation problem optimally. The algorithm in the following aims to utilize only those $N \ll 2^K$ scenarios that have the strongest impact on overall cell load levels. The algorithm is designed in such a way that the discrete interference level used as an approximation is always an over-estimator of the actual interference plus noise. This is achieved by adding in the reformulated

problem the following constraint:

$$\sum_{j\in\{\mathcal{C}\setminus\{k\}\}}\sum_{q=1}^{Q} B_{qk}B_{qj}p_j g_{jm} + \epsilon \leq \sum_{n=1}^{N}\phi_{nkm}\Psi_{nkm}\ \forall k,m \tag{5.5}$$

where $\phi_{nkm} = 1$ if interference scenario n is used as an approximation for the link between DP m and cell k, and $\phi_{nkm} = 0$ otherwise. Since exactly one interference scenario applies for each pair of cell k and DP m, it has to hold that

$$\sum_{n}\phi_{nkm} = 1\ \forall k,m. \tag{5.6}$$

The elements ϕ_{nkm} are arranged in the three-dimensional binary array $\boldsymbol{\phi} \in \{0,1\}^{N\times K\times M}$. The proposed inner approximation of Eq. (5.4h) is the following:

$$\sum_{m=1}^{M}\sum_{n=1}^{N} D_m\phi_{nkm}A_{km}\zeta^{+}_{\tau^{\mathrm{MIN}}}\left(\frac{p_k g_{km}}{\Psi_{nkm}}\right) \leq \eta_k^{\mathrm{K}}\sum_{q=1}^{Q}\left(\eta_q^{\mathrm{Q}}B_{qk}w_q\right)\ \forall k \tag{5.7}$$

The term $\zeta^{+}_{\tau^{\mathrm{MIN}}}\left(p_k g_{km}/\Psi_{nkm}\right)$ in Eq. (5.7) can be pre-computed for all combinations of (n,k,m). Approximating the constraints (5.4h) with (5.5) and (5.7) leads to an upper bound approximation of the interference and accordingly the required bandwidth for each user. Solutions obtained from using the latter set of constraints are therefore feasible for the original problem. A possible selection of discrete interference levels The constraints (5.4g) are approximated correspondingly with

$$A_{km}p_k^{\mathrm{S}}g_{km} \geq \gamma_i^{\mathrm{MIN}}\sum_{n=1}^{N}\phi_{nkm}A_{km}\Psi_{nkm}\ \forall\ k, \left\{m,q|\sum_{k}A_{km}B_{qk} = 1\right\}. \tag{5.8}$$

To enable the correct parameter selection for which Eq. (5.8) has to hold, it is modified using a big-M reformulation [CPP13] where a term $\xi(1-\sum_{k=1}^{K}A_{km}B_{qk})$ is subtracted on the right-hand side to ensure that the SINR-constraint is always fulfilled if $\sum_{k=1}^{K}A_{km}B_{qk} = 0$, i.e. when DP m is not serviced in slice q:

$$A_{km}p_k^{\mathrm{S}}g_{km}A_{km}p_k^{\mathrm{S}}g_{km} \geq \gamma_q^{\mathrm{MIN}}\left(\sum_{n=1}^{N}\phi_{nkm}A_{km}\Psi_{nkm} - \xi\left(1-\sum_{k=1}^{K}A_{km}B_{qk}\right)\right)\ \forall\ k,q,m \tag{5.9}$$

For this property to hold, the parameter ξ needs to be chosen such that $\xi \geq \max_{n,k,m}\Psi_{nkm}$. The reformulated problem (5.4) reads as follows:

$$\underset{Z,\boldsymbol{w},\boldsymbol{A},\boldsymbol{B},\boldsymbol{\phi}}{\text{maximize}}\quad Z \tag{5.10a}$$

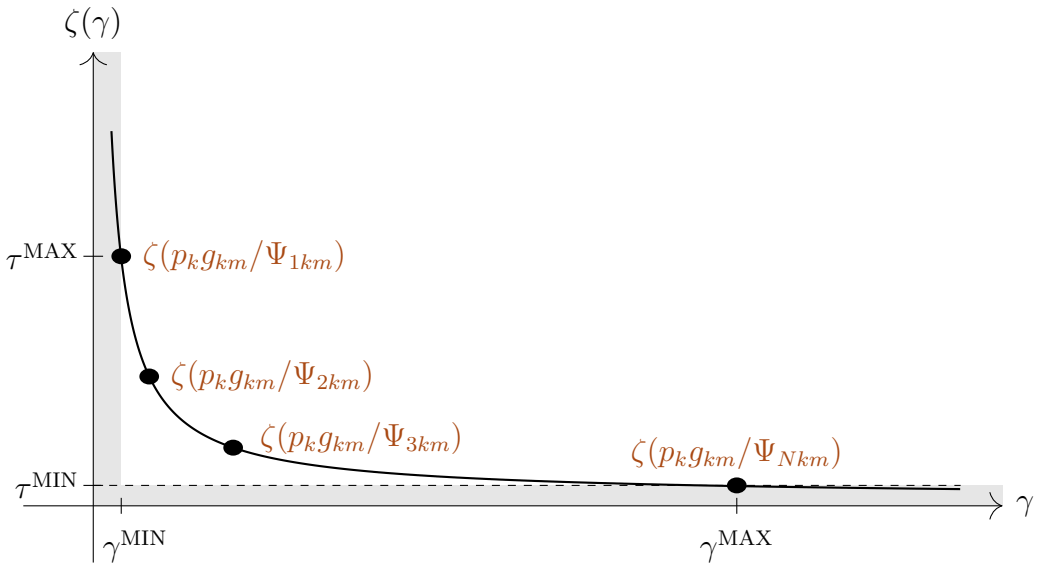

Figure 5.1. Illustration of the load function for an example of discrete interference terms. The density of the discretization is chosen to be higher for SINR-levels where the load function $\zeta(\gamma)$ has a steeper slope.

$$\begin{aligned} \text{subject to} \quad & (5.4\text{b}) - (5.4\text{d}), (5.4\text{e}), (5.5), (5.6), (5.7), (5.9) \\ & Z, w_q \in \mathbb{R}_{0+} \; \forall \; q && (5.10\text{b}) \\ & A_{km}, B_{qk}, \phi_{nkm} \in \{0,1\} \; \forall \; q, k, m, n && (5.10\text{c}) \end{aligned}$$

The resulting optimization problem (5.10) is still nonlinear because of multiple bilinear products of two optimization variables in the constraints, specifically $A_{km}B_{qk}$ in Eq. (5.4g) and Eq. (5.4e), $\phi_{nkm}A_{km}$ in Eq. (5.9) and Eq. (5.7), $B_{qk}B_{qj}$ in Eq. (5.5) and $B_{qk}w_q$ in Eq. (5.7). These bilinear terms are reformulated using the constraint sets $\mathcal{B}$ and $\mathcal{L}$ as defined in Sec. 3.2.2.

To replace the aforementioned bilinear products, the lifting strategy introduced in Sec. 3.2.2 is used that linearizes the problem at the cost of increased dimensionality. Denote as

$$w^{\mathrm{B}}_{qk} = B_{qk} w_q \; \forall \; q, k \tag{5.11}$$

the resources cell k utilizes in slice q. Because the maximum amount of resources utilized by any cell in any slice is nonnegative and bounded by $w_q \leq \overline{w}$, Eq. (5.11) is reformulated based on Eq. (3.3) to the following inequalities:

$$\begin{aligned} & \forall \; q, k : && (5.12\text{a}) \\ & \quad w^{\mathrm{B}}_{qk} \geq w_q - (1 - B_{qk})\, \overline{w} && (5.12\text{b}) \end{aligned}$$

$$w_{qk}^{\mathrm{B}} \leq w_q \tag{5.12c}$$
$$w_{qk}^{\mathrm{B}} \leq \overline{w} B_{qk} \tag{5.12d}$$

A real matrix $\boldsymbol{w}^{\mathbf{B}} \in \mathbb{R}_{0+}^{Q \times K}$ shall in the following represent all arranged elements w_{qk}^{B}. To write the set of linear inequalities (5.12) in a shorter form, the set $\mathcal{L}$ introduced in Sec. 3.2.2 will be used in the following.
The term $\sum_{j \in \{\mathcal{C} \setminus \{k\}\}} \sum_{q=1}^{Q} B_{qk} B_{qj}$ in Eq. (5.5) contains the product $B_{qk} B_{qj}$ of two binary parameters, which is also reformulated based on the lifting strategy introduced in Sec. 3.2.2. Denote as

$$B_{qkj}^{\mathrm{INT}} = \begin{cases} B_{qk} B_{qj} \; \forall \; q, k, j : k \neq j \\ 0 \; \forall \; q, k, j : k = j \end{cases} \tag{5.13}$$

an auxiliary parameter indicating whether cells k and j are interfering with each other in slice q. The bilinear product $B_{qk} B_{qj}$ computing the non-zero B_{qkj}^{INT}, for which $k \neq j$, can be recast into an equivalent formulation with the following inequalities:

$$\forall \; q, k \neq j : \tag{5.14a}$$
$$B_{qkj}^{\mathrm{INT}} \leq B_{qk} \tag{5.14b}$$
$$B_{qkj}^{\mathrm{INT}} \leq B_{qj} \tag{5.14c}$$
$$B_{qkj}^{\mathrm{INT}} \geq B_{qk} + B_{qj} - 1 \tag{5.14d}$$

The elements B_{qkj}^{INT} are arranged in the three-dimensional binary array $\boldsymbol{B}^{\mathbf{INT}} \in \{0,1\}^{Q \times K \times K}$, and the constraints (5.14) is written in the following using the set $\mathcal{B}$ introduced in Sec. 3.2.2.
Further denote the auxiliary optimization parameters $A_{ikm}^{\mathrm{B}} \triangleq A_{km} B_{qk}$, $\phi_{nkm}^{\mathrm{A}} \triangleq \phi_{nkm} A_{km}$, with their elements arranged in the three-dimensional binary arrays $\boldsymbol{A}^{\mathbf{B}} \in \{0,1\}^{Q \times K \times M}$ and $\boldsymbol{\phi}^{\mathbf{A}} \in \{0,1\}^{N \times K \times M}$, respectively. To ensure that these auxiliary parameters are equal to the bilinear products of the original optimization parameters again the linear inequality constraints in $\mathcal{B}$ are utilized. Implementing the reformulation of bilinear terms in the optimization problem results in the following MILP:

$$\underset{Z, \boldsymbol{w}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A}^{\mathbf{B}}, \boldsymbol{B}^{\mathbf{INT}}, \boldsymbol{\phi}, \boldsymbol{\phi}^{\mathbf{A}}, \boldsymbol{w}^{\mathbf{B}}}{\text{maximize}} \quad Z \tag{5.15a}$$

$$\text{subject to} \quad (5.4\text{b}) - (5.4\text{d})$$
$$\sum_{k=1}^{K} A_{ikm}^{\mathrm{B}} \leq S_{qm} \; \forall \; q, m \tag{5.15b}$$

$$\gamma_q^{\mathrm{MIN}} \left(\sum_{n=1}^{N} \phi_{nkm}^{\mathrm{A}} \Psi_{nkm} - \xi \left(1 - \sum_{k=1}^{K} A_{qkm}^{\mathrm{B}} \right) \right) \leq A_{km} p_k^{\mathrm{S}} g_{km} \; \forall \; k, q, m \tag{5.15c}$$

$$\sum_{m=1}^{M} \sum_{n=1}^{N} D_m \phi_{nkm}^{\mathrm{A}} \zeta_{\tau^{\mathrm{MIN}}}^{+} \left(\frac{p_k g_{km}}{\Psi_{nkm}} \right) \leq \eta_k^{\mathrm{K}} \sum_{q=1}^{I} \left(\eta_q^{\mathrm{Q}} w_{qk}^{\mathrm{B}} \right) \; \forall \; k \tag{5.15d}$$

$$\sum_{q=1}^{I} \sum_{j=1}^{K} B_{qkj}^{\mathrm{INT}} p_j^{\mathrm{S}} g_{jm} + \epsilon \leq \sum_{n=1}^{N} \phi_{nkm} \Psi_{nkm} \; \forall \; k, m \tag{5.15e}$$

$$\sum_{n} \phi_{nkm} = 1 \; \forall \; k, m \tag{5.15f}$$

$$\{ w_q, \overline{w}, B_{qk}, w_{ik}^{\mathrm{B}} \} \; \in \; \mathcal{L} \; \forall \; q, k \tag{5.15g}$$

$$\{ \phi_{nkm}, A_{km}, \phi_{nkm}^{\mathrm{A}} \} \; \in \; \mathcal{B} \; \forall \; n, k, m \tag{5.15h}$$

$$\{ A_{km}, B_{ik}, A_{ikm}^{\mathrm{B}} \} \; \in \; \mathcal{B} \; \forall \; q, k, m \tag{5.15i}$$

$$\{ B_{qk}, B_{qj}, B_{qkj}^{\mathrm{INT}} \} \; \in \; \mathcal{B} \; \forall \; q, j \neq k; B_{qkj}^{\mathrm{INT}} = 0 \; \forall \; q, j = k \tag{5.15j}$$

$$Z, w_q, w_{qk}^{\mathrm{B}} \in \mathbb{R}_{0+} \; \forall \; q, k \tag{5.15k}$$

$$A_{km}, B_{qk}, \phi_{nkm}, \phi_{nkm}^{\mathrm{A}}, B_{qkj}^{\mathrm{INT}}, A_{ikm}^{\mathrm{B}} \in \{0, 1\} \; \forall \; q, k, j, m, n \tag{5.15l}$$

The problem formulated in (5.15) is linear in all optimization variables and therefore can be solved using conventional MILP solvers.

5.4 Simulation Results

A mobile communication network is simulated with the parameters outlined in Table 5.1. The network, as shown in Fig. 5.2 contains three macro cells and six small cells that are located along the edges of the macro cell coverage areas. The simulation considers three possible configurations for the interference scenario, corresponding respectively to a full interference setting, a removal to the strongest interfering cell and the removal of the two strongest interfering cells. The final interference level corresponds to the no-interference case. Specifically, the discrete interference levels Ψ_{nkm} with $N = 4$ are computed as

$$\Psi_{1km} = \sum_{j \in \{\mathcal{C} \setminus k\}} p_j g_{jm} + \lambda \overline{w}, \tag{5.16}$$

$$\Psi_{2km} = \sum_{j \in \{\mathcal{C} \setminus k, \kappa_{km}^{\mathrm{P}}\}} p_j g_{jm} + \lambda \overline{w}, \tag{5.17}$$

$$\Psi_{3km} = \sum_{j \in \{\mathcal{C} \setminus k, \kappa_{km}^{\mathrm{P}}, \kappa_{km}^{\mathrm{S}}\}} p_j g_{jm} + \lambda \overline{w}. \tag{5.18}$$

Table 5.1. Simulation parameters of a downlink LTE network for resource efficiency minimization. The resource and DP allocation is optimized. Resource consumption performance is averaged over 500 simulations with fixed base station positions and randomly distributed DPs.

Area size	1000×1000 m
Noise power spectral density λ	-145 dBm/Hz
Max. time-frequency resources $\overline{w}$	20 MHz
Number of DPs M	40
Position of macro BS	MBS1 at [100m, 800m] MBS2 at [900m, 800m] MBS3 at [500m, 140m]
MBS transmit power spectral density p^{S}	-27 dBm/Hz
MBS antenna gain $\tilde{g}^{\mathrm{ABS}}$	15dB
Position of pico BS	PBS1 at [500m, 900m] PBS2 at [500m, 600m] PBS3 at [300m, 400m] PBS4 at [150m, 200m] PBS5 at [700m, 400m] PBS6 at [850m, 200m]
PBS transmit power spectral density p^{S}	-37 dBm/Hz
PBS antenna gain $\tilde{g}^{\mathrm{ABS}}$	5dB
DP antenna gain $\tilde{g}^{\mathrm{ADP}}$	0dB
Propagation loss $\tilde{g}^{\mathrm{PATH}}$	3GPP TS 36.814 [3GP16]
Cell bandwidth efficiency η_k^{K}	1
Slice bandwidth efficiency η_k^{K}	1 (normal slice) 0.5 (reliability slice)
SINR requirement γ_q^{MIN}	-7dB (normal slice) 0dB (reliability slice)
SINR threshold γ^{MAX}	20dB

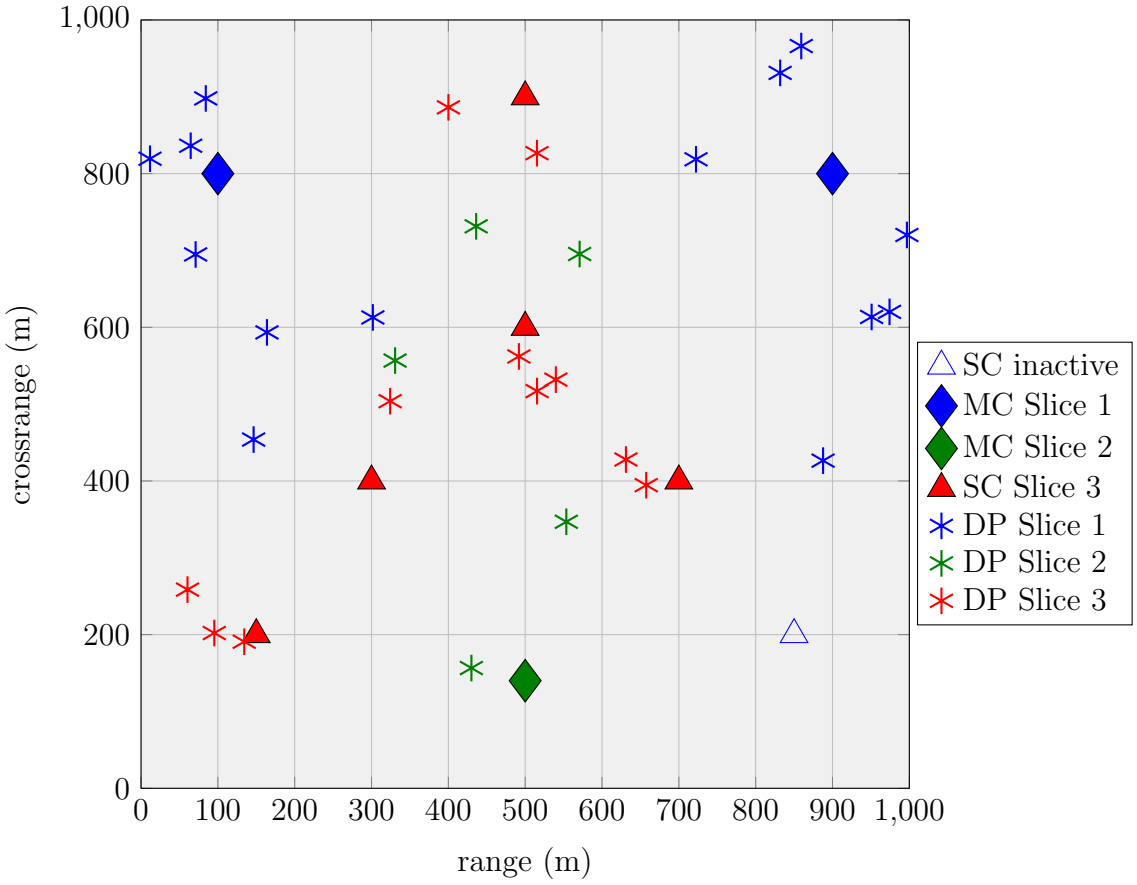

Figure 5.2. Illustration of a typical resource distribution, slices, and user allocation result. All SCs are automatically allocated to a separate resource slice from the ones utilized by the macro cells.

The final interference scenario models the noise-only case $\Psi_{4km} = \lambda\overline{w}$. The optimization problem in (5.15) is solved using CVX for MATLAB [GB14, GB08] and Gurobi as a MILP solver [GUR]. For each network scenario, $M = 40$ demand points are randomly distributed in the simulated area. The network scenarios described in this section have been solved on a standard workstation with an Intel i7-7600 processor, with a solver time of approximately one minute for each scenario. A typical result of the proposed method is illustrated in Fig. 5.2. Three orthogonal resource slices, indicated by color in Fig. 5.2, are utilized ($Q = 3$). The proposed method automatically allocates all SCs to a separate slice from those utilized by the MCs. One SC in the bottom right corner of the map is not utilized at all, resulting from a lack of DPs in close proximity. Different services envisioned for 5G, as discussed in Sec. 1.1, might require the use of transmission and coding schemes that require better SINR-levels and that sacrifice bandwidth efficiency for reliability. The corresponding parameters for this “reliability”

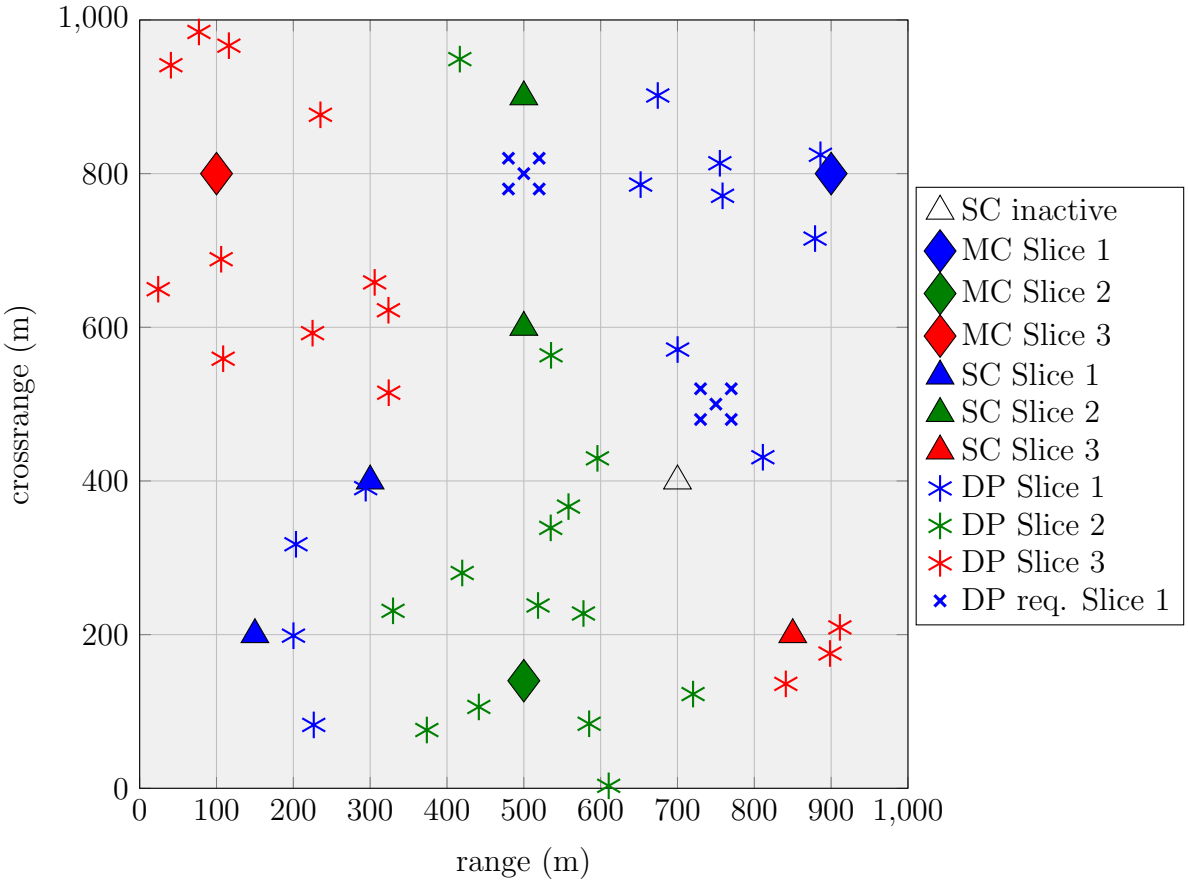

Figure 5.3. Illustration of the resource slicing distribution with one reliability-focused slice. Two clusters of five DPs each are specifically requesting the reliability slice.

slice are shown in Table 5.1. In the example scenario shown in Fig. 5.3, two clusters were added with five DPs each that explicitly request service from slice $i = 1$. This slice is set up as a "high reliability" slice that could provide for example coverage to machine-to-machine services. Because of these requirements, the transmission scheme used in this slice has lower bandwidth efficiency ($\eta_1^{\mathrm{I}} = 0.5$) and requires good SINR ($\gamma_1^{\mathrm{MIN}} = 0\mathrm{dB}$). As observable, the macro cell in the upper right corner of the map is reserved almost exclusively for the service of the DPs requesting Slice 1. Interference to these DPs is also actively regulated because only two SCs in the lower right region of the map also share this "high reliability" slice. The algorithm also chooses to service some other users in this slice, even though they had not specifically requested it.

In the following the effect of cell planning with multiple orthogonal resource pools on the resource efficiency of the system, as optimized by solving problem (5.15), is analyzed. As a baseline method to optimize the resource efficiency the state-of-the-art approach of full frequency reuse, and an allocation of DPs to the cell providing the

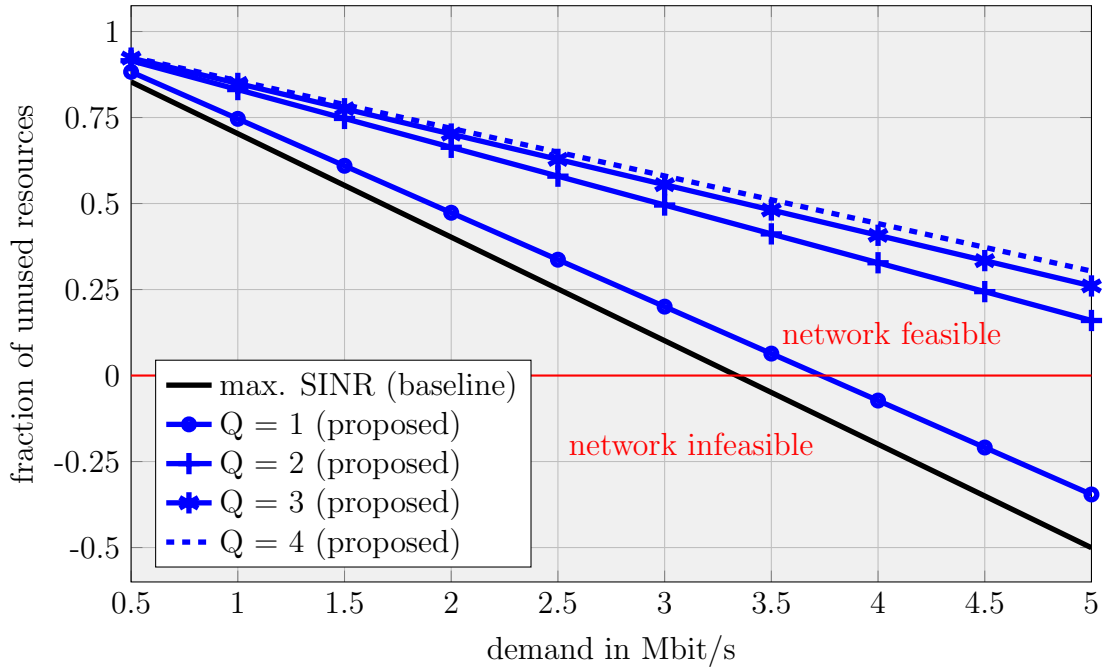

Figure 5.4. Network resource utilization of the proposed resource slicing optimization for varying user demand (simulation 1). The proposed method benefits from the availability of additional, orthogonally operating resource slices.

strongest signal is used. To allow for a comparison with this approach, the slices in the proposed method are modeled with equal parameters as outlined in Table 5.1.

In a simulation to evaluate the resource efficiency of the proposed method, 250 network scenarios with the DP demand in each scenario increasing from $D_m = 0.5$ Mbit/s to $D_m = 5$ Mbit/s, and $M = 40$ randomly placed DPs for each scenario, where the resulting levels of unused resources Z are averaged over all scenarios. As observable in Fig.5.4, the proposed method with $Q = 1$ yields a resource efficiency slightly higher than the baseline method, but remark that both methods require more resources than are actually available for high demand, which means that Z is negative. For $Q = 2$, the resource efficiency vastly improves, and increasing the maximum number of slices to $Q = 3$ and $Q = 4$ improves the performance even further, but shows diminishing returns.

In a second simulation, the network for this scenario is simplified from that shown in Fig. 5.2 such that only the three macro cells and the small cell in the center are deployed. The DPs are randomly placed in the simulated area using a uniform probability distribution. As observable in Fig. 5.5, the proposed method with $Q = 1$ yields a resource efficiency slightly higher than the baseline method, but remark that both methods require more resources than are actually available for high demand, which means that Z

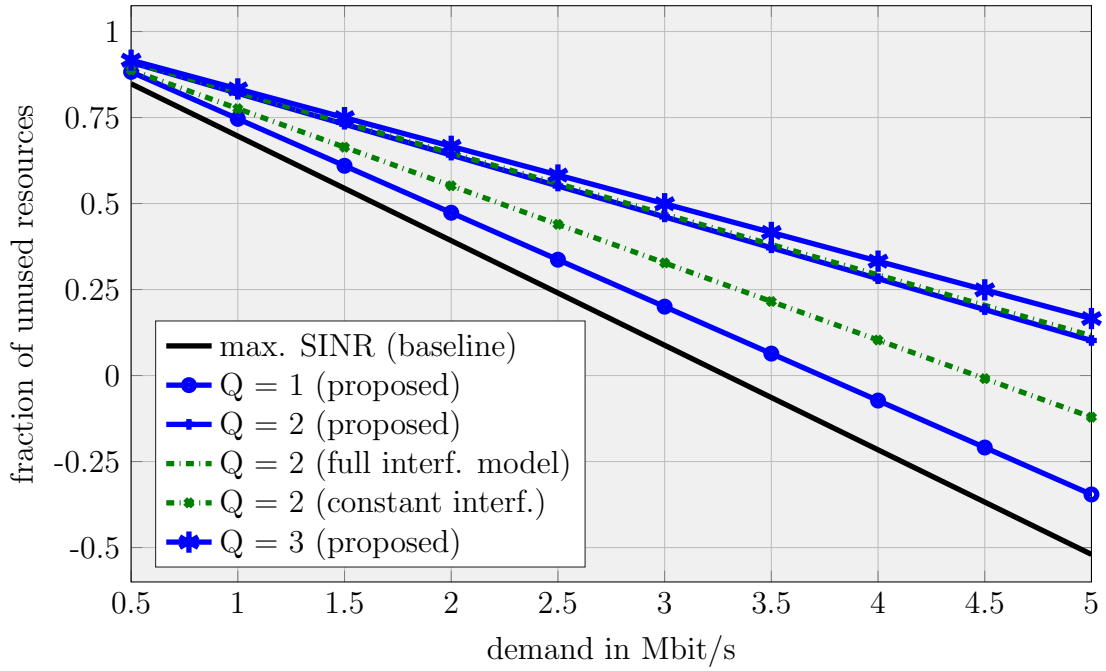

Figure 5.5. Resource consumption comparison of the proposed resource slicing method for decreased network size (simulation 2). The proposed scheme achieves better performance due to its adaptive interference modeling compared to the established static interference modeling.

is negative. For $Q = 2$, the remaining resources are also shown for modified versions of the algorithm. In the first modified algorithm, indicated as "full interference model", it is shown that selecting Ψ_{nkm} such that all possible combinations of active cells are considered only provides marginal performance increases. The second modified algorithm assumes a constant interference model used for example in [CBdVCP17] and the references therein, which shows significantly decreased resource efficiency due to the worse approximation of the actual interference.

In a third numerical experiment, 500 network scenarios are simulated, where the 6 small cells are one by one activated in a location randomly chosen from the 6 available locations shown in Fig. 5.2. This simulation is designed to evaluate the benefits of network densification, represented by the number of additionally deployed small cells. The results are shown in Fig. 5.6. As observable, the resource efficiency for the max. SINR method overall shows no real benefit from densifying the network. Even worse, the resource consumption first decreases if The proposed method however already shows some gains for $I = 1$, where only user allocation and cell activity status is optimized. It is observable that especially two deployed small cells appear to be an unsuitable solution for the given network, with both the baseline method and the proposed method

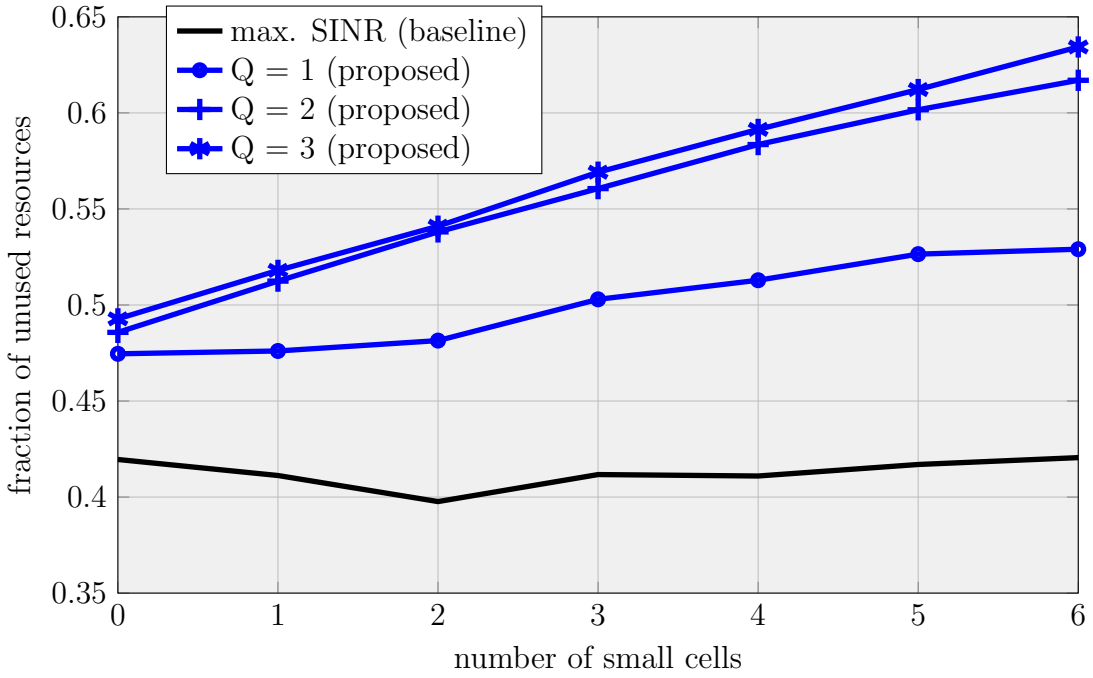

Figure 5.6. Resource consumption comparison for varying number of small cells (simulation 3). Reliable performance gains are only achievable if the proposed method is used with multiple resource slices.

with $Q = 1$ showing a decrease in performance. If additionally $Q = 2$ or $Q = 3$ slices operating on orthogonal resource pools are allowed, the proposed method demonstrates gains in resource efficiency and the benefits of network densification.

5.5 Summary

A method was introduced to optimize the resource allocation in a heterogeneous wireless communication network. The proposed scheme relies on minimizing the resource consumption of the network by solving a joint optimization of the allocation of cells to orthogonally operating slices of resources, the dimensioning of these slices, and the allocation of DPs to cells. An adaptive estimation of the interference levels is used in the optimization which significantly improves upon the static interference model used in common literature. The optimization is carried out such that specific slices can be set up with different QOS requirements, such as SINR constraints or modulation and coding schemes with different bandwidth efficiencies, to account for the service-centric network design paradigm for 5G discussed in Sec. 1.1. The proposed method

achieves superior resource efficiency compared to the baseline approach, and enables further performance gains through network densification when additional small cells are deployed.

Chapter 6

Energy Consumption Minimization

6.1 Introduction and Contributions

In dense and heterogeneous wireless networks, the existing cell architecture is supplemented with additional cells containing base stations of variable size, both in transmit power and coverage area. This densification of the network has been identified as a promising approach for the next decades of wireless communications. The scalability of such networks, especially with regards to network energy consumption, has come under recent scrutiny [BLM+14, GTM+16, WWH+17, ZSB+16]. Due to the increase in intercell interferences limiting the achievable data throughput, novel control schemes for such networks need to be devised that supersede the established strategy of deploying additional cells without increasing the amount of coordination between them [AZDG16, CSS+14, IRH+14]. The wireless communication networks of the future are envisioned to have a significantly higher energy efficiency in terms of energy consumption per transmitted bit of data. In the 5G standard, this will be achieved trough intelligent switching of each cell's operation between active phases and sleep modes - abandoning the always-on and always-connected concept of contemporary base stations - a dynamic scaling of the transmit power, and an energy-focused design of multi-antenna systems [LKB+14, CSS+14, CZB+10, VHD+11].

In this chapter, a method is proposed for minimizing the energy consumption of the wireless communication network, subject to cell load constraints that prevent cells from being unable to serve the demand of associated users with their available time-frequency resources. This approach is suitable for the planning of the network parameters ahead of operation, and complements energy efficient transceiver techniques commonly applied in-operation for example to maximize instantaneous data rates.

6.1.1 State-of-the-Art

In previous research, extensive effort has been invested into the analysis and optimization of cell loads for heterogeneous mobile communication networks [YY17, SY12b,

SY13, YPC+15]. The cell load has been used in various schemes to optimize the transmit powers [HYLS15, YLY16], in the design of energy-efficient beamformers for multi-antenna systems [CZL16, MHLT11], and to optimize the cell on-off status to enable scheduling for sleep mode and activity periods [CG17, LYHS15]. These methods share one fundamental disadvantage, which is that they cannot jointly optimize the transmit power and the cell activity status. Switching cells off is just considered implicitly, as the transmit power being scaled down to zero [HYLS15, KU16]. The transmit power in a practical system however might be lower-bounded by a nonzero level, for example due to transmit power independent losses and nonlinearities in the power amplifiers [KB02, DDG+12, ARFB10]. Heuristic approaches also heavily rely on the cell load being a strictly decreasing continuous function of its transmit power, which requires multiple simplifications in the way how the network is modeled, particularly regarding the used adaptive modulation and coding schemes. For example, the load a user adds to a cell needs to be a strictly decreasing function of the user's SINR, the assignment of users to cells needs to be constant and the operable transmit power range needs to be lower-bound by zero, which all do not necessarily apply in practical systems.

6.1.2 Contributions and Overview

The proposed approach for energy consumption minimization expands upon state of the art solutions in the following aspects: The transmit power and the activity status of the cell (on or off) are jointly optimized. This leads to a mixed-integer problem as, e.g. the transmit power is optimized on a continuous scale and the cell activity indicator is binary. While the original optimization problem is nonlinear and computationally intractable to solve, a linear inner approximation is proposed. The solution of this approximate problem is always feasible for the original problem. The proposed method easily incorporates additional convex constraints such as minimum transmit power and minimum SINR threshold constraints as well as upper bounds on the user rates due to finite modulation and coding schemes. The assignment of users to cells is one of the design parameters, and changes dynamically according to which cell provides the strongest signal. The proposed scheme also allows the incorporation of other user allocation rules.

The remainder of this chapter is structured as follows: In Section 6.2 the system model from Chapter 2 is expanded for the wireless communication network and provide a formulation of the energy minimization problem. A mixed-integer nonlinear programming (MINLP) approach to minimize the network's energy consumption is discussed in Section 6.3, for which an inner linear approximation (MILP) is provided. Simulation

results for different energy consumption models and a comparative analysis between the proposed and alternative methods are provided in Section 6.4. Finally, the results are summarized in 6.5.

6.2 Problem Formulation

Expanding upon the system model introduced in Chapter 2, the transmit power p_k will in the following be considered as an optimization parameter. In practical networks, p_k is generally confined to lie in a the interval

$$0 < P_k^{\mathrm{MIN}} \leq p_k \leq P_k^{\mathrm{MAX}}, \tag{6.1}$$

where, due to physical hardware limitations, such as linearity constraints in the power amplifiers and radiation efficiency requirements of the antenna the thresholds P_k^{MIN} and P_k^{MAX} are positive (excluding $P_k^{\mathrm{MIN}} = 0$) and finite [KB02, DDG+12, ARFB10]. Let $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]^T$ be the vector of all transmit powers.
The on-off activity status of cells is denoted with the binary model parameter

$$z_k = \begin{cases} 1 & \text{if cell } k \text{ is active} \\ 0 & \text{otherwise} \end{cases} . \tag{6.2}$$

and the vector $\boldsymbol{z} = [z_1, z_2, \ldots, z_K]^\top$ representing the activity status of all cells in the network.
The energy consumption of a cell shall be defined as

$$E_k^{(\Gamma)} = \Gamma\left(z_k, p_k, \rho_k\right) \tag{6.3}$$

where $\Gamma\left(z_k, p_k, \rho_k\right)$ is an arbitrary linearly increasing function of the cell's on-off status z_k, transmit power p_k and load ρ_k. For example, the energy consumption function used in Eq. (6.3) can be defined as

$$\Gamma\left(z_k, \tilde{p}_k, \rho_k\right) = T_0 P_k^{\mathrm{MAX}} \left(\kappa_1 z_k + \kappa_2 \frac{\tilde{p}_k}{P_k^{\mathrm{MAX}}} + \kappa_3 \tilde{\rho}_k\right) \tag{6.4}$$

where the parameters κ_1, κ_2 and κ_3 are weighting factors for the cell's energy consumption based on the on-off status, transmit power, and load factor, respectively, and T_0 is a time constant. The load factor of a cell can impact its power consumption because it reflects the amount of its utilization [SKYK11]. Recent network models therefore have

established that, especially for small cells, the power consumption is best modeled as a function of the cell load in addition to the transmit power [KU16, YLY16]. Note that the terms z_k, $\tilde{p}_k/P_k^{\mathrm{MAX}}$ and $\tilde{\rho}_k$ cannot exceed the value 1, for each cell k. For more sophisticated models for the power consumption of mobile communication BSs, which incorporate energy consumption of wired backhaul, and individual factors for all components of the BS, refer to [HBB11, BC11, DDG+12, DJM14]. Since the propsed model can use any combination of the three factors in Eq. (6.3), a highly flexible approach for energy minimization is obtained.

In this section, an optimization problem in the form of a MIP is formulated to minimize the energy consumption of the wireless network as defined in Eq. (6.3), subject to DP to BS allocation-, minimum SINR- and cell load constraints. The system model is based on Secs. 2.3 and 2.2. Using as optimization parameters the binary cell activity indicator $\boldsymbol{z} \in \{0,1\}^{K\times 1}$ and allocation indicator $\boldsymbol{A} \in \{0,1\}^{M\times K}$, the continuous transmit power parameter $\boldsymbol{p} \in \mathbb{R}_{0+}^{K\times 1}$ and the cell load $\boldsymbol{\rho} \in \mathbb{R}_{0+}^{K\times 1}$, the energy minimization problem can be formulated as following:

$$
\begin{aligned}
\underset{\boldsymbol{z},\boldsymbol{p},\boldsymbol{A},\boldsymbol{\rho}}{\text{minimize}} \quad & \sum_{k=1}^{K} \Gamma\left(z_k, p_k, \rho_k\right) && \text{(6.5a)}\\
\text{subject to} \quad & P_k^{\mathrm{MIN}} \leq p_k \leq P_k^{\mathrm{MAX}} \; \forall \; k && \text{(6.5b)}\\
& \sum_{k=1}^{K} A_{km} = 1 \; \forall \; m && \text{(6.5c)}\\
& \sum_{k=1}^{K} A_{km} \leq z_k \; \forall \; k, m && \text{(6.5d)}\\
& \sum_{k=1}^{K} A_{km}\theta_k p_k g_{km} \geq z_j \theta_j p_j g_{jm} \; \forall \; j, m && \text{(6.5e)}\\
& \sum_{k=1}^{K} A_{km} p_k g_{km} - \gamma^{\mathrm{MIN}} \left(\sum_{j} z_j (1 - A_{jm}) p_j g_{jm} + \sigma^2 \right) \geq 0 \; \forall \; m && \text{(6.5f)}\\
& \rho_k = \sum_{m=1}^{M} A_{km} \frac{d_m}{W \eta_{km}^{\mathrm{BW}}} \zeta_{\tau^{\mathrm{MIN}}}^{+} \left(\frac{p_k g_{km}}{\sum_{j=1}^{K} z_j (1 - A_{jm}) p_j g_{jm} + \sigma^2} \right) \; \forall \; k && \text{(6.5g)}\\
& \rho_k \leq 1 \; \forall \; k && \text{(6.5h)}\\
& z_k, A_{km} \in \{0,1\} \; \forall \; k, m && \text{(6.5i)}\\
& p_k \in \mathbb{R}_{0+} \; \forall \; k && \text{(6.5j)}
\end{aligned}
$$

In problem (6.5), the objective (6.5a) aims to minimize the overall systems' energy consumption, which is the sum of the energy consumption of individual cells as defined in (6.3) and (6.4). The constraint (6.5b) defines the feasible transmit power range of

cell k restricted according to (6.1). The fixed data rate demand of each DP m is served by exactly one cell k, and only active cells $\{k|z_k = 1\}$ can serve any DP, as specified by (6.5c) and (6.5d), respectively. Constraint (6.5e) enforces that, each DP m is allocated to the cell k that provides highest product of received signal power and bias value. The load constraint that cell k has to satisfy, as defined in (2.10), is specified in (6.5h). Problem (6.5) is a combinatorial and nonconvex MINLP, and thus generally very difficult to solve. As discussed in Sec. 3.2.1, it is therefore advisable to find an MILP that represents a linear inner approximation or a linear reformulation of the original MINLP, which will be discussed in the following Sec. 6.3.

6.3 Energy Minimization Scheme

The objective function (6.5a) and constraints (6.5e), (6.5f) and (6.5h) contain the bilinear term $z_k p_k$. A new variable $\tilde{p}_k \triangleq p_k z_k$ is introduced to reformulate (6.5) as the following equivalent problem:

$$\underset{\boldsymbol{z},\tilde{\boldsymbol{p}},\boldsymbol{A},\boldsymbol{\rho}}{\text{minimize}} \quad \sum_{k=1}^{K} \Gamma\left(z_k, \tilde{p}_k, \rho_k\right) \tag{6.6a}$$

$$\text{subject to} \quad z_k P_k^{\text{MIN}} \leq \tilde{p}_k \leq z_k P_k^{\text{MAX}} \; \forall \; k \tag{6.6b}$$

$$(6.5\text{c}) - (6.5\text{d})$$

$$\sum_{k=1}^{K} A_{km} \theta_k \tilde{p}_k g_{km} \geq \theta_j \tilde{p}_j g_{jm} \; \forall \; j, m \tag{6.6c}$$

$$\sum_{k=1}^{K} A_{km} \tilde{p}_k g_{km} - \gamma^{\text{MIN}} \left(\sum_{j} (1 - A_{jm}) \tilde{p}_j g_{jm} + \sigma^2 \right) \geq 0 \; \forall \; m \tag{6.6d}$$

$$\rho_k = \sum_{m=1}^{M} A_{km} \frac{d_m}{W \eta_{km}^{\text{BW}}} \zeta_{\tau^{\text{MIN}}}^{+} \left(\frac{\tilde{p}_k g_{km}}{\sum_{j=1}^{K} (1 - A_{jm}) \tilde{p}_j g_{jm} + \sigma^2} \right) \; \forall \; k \tag{6.6e}$$

$$\rho_k \leq 1 \; \forall \; k \tag{6.6f}$$

$$z_k, A_{km} \in \{0, 1\} \; \forall \; k, m \tag{6.6g}$$

$$\tilde{p}_k \in \mathbb{R}_{0+} \; \forall \; k \tag{6.6h}$$

Using a lifting strategy, auxiliary parameters will in the following be introduced to represent bilinear products of optimization variables, which yields a more tractable, linear problem structure at the cost of increased problem dimensionality. Towards this aim, the bilinear products of binary allocation parameters A_{km} and cell transmit powers $\tilde{p}_k$ in Eqs. (6.6c), (6.6d) and (6.6e) have to be linearized using the procedure

outlined in Sec. 3.2.2.

Let Ω_{km} be an auxiliary optimization parameter and denote the corresponding matrix $\boldsymbol{\Omega} \in \mathbb{R}_{0+}^{K\times M}$. For the proposed lifting approach, $\left(\tilde{p}_k, P_k^{\mathrm{MAX}}, A_{km}, \Omega_{km}\right) \in \mathcal{L}\ \forall k, m$ is installed in problem (6.6), which enforces that $\Omega_{km} = \tilde{p}_k A_{km}$, such that (6.5) can be reformulated as:

$$\begin{aligned}
\underset{\boldsymbol{z},\boldsymbol{p},\boldsymbol{A},\boldsymbol{\rho},\boldsymbol{\Omega}}{\text{minimize}} \quad & \sum_{k=1}^{K} \Gamma\left(z_k, \tilde{p}_k, \rho_k\right) && \text{(6.7a)}\\
\text{subject to} \quad & (6.5\text{c}) - (6.5\text{d}), (6.6\text{b}), (3.4) && \\
& \sum_{k=1}^{K} \Omega_{km}\theta_k g_{km} \geq z_j \theta_j \tilde{p}_j g_{jm}\ \forall\, j, m && \text{(6.7b)}\\
& \sum_{k=1}^{K} \Omega_{km} g_{km} - \gamma^{\mathrm{MIN}} \left(\sum_j (1-\Omega_{jm}) g_{jm} + \sigma^2 \right) \geq 0\ \forall\, m && \text{(6.7c)}\\
& \rho_k = \sum_{m=1}^{M} A_{km} \frac{d_m}{W \eta_{km}^{\mathrm{BW}}} \zeta_{\tau^{\mathrm{MIN}}}^{+} \left(\frac{\tilde{p}_k g_{km}}{\sum_{j=1,\ldots,K} (\tilde{p}_j - \Omega_{jm}) g_{jm} + \sigma^2} \right)\ \forall\, k && \text{(6.7d)}\\
& \rho_k \leq 1\ \forall\, k && \text{(6.7e)}\\
& \left(\tilde{p}_k, P_k^{\mathrm{MAX}}, A_{km}, \Omega_{km}\right) \in \mathcal{L}\ \forall\, k, m && \text{(6.7f)}\\
& z_k, A_{km} \in \{0,1\}\ \forall\, k, m && \text{(6.7g)}\\
& \tilde{p}_k, \Omega_{km} \in \mathbb{R}_{0+}\ \forall\, k && \text{(6.7h)}
\end{aligned}$$

From (6.5) to (6.7), the auxiliary parameter $\boldsymbol{\Omega}$ has been used in constraints (6.7b), (6.7c) and (6.7d) to replace $\Omega_{km} = \tilde{p}_k A_{km}$, whereas the remaining optimization parameters remain unchanged. The solution of problem (6.7) can therefore be used to easily obtain the corresponding solutions for problem (6.5) and vice-versa. Thus, both formulations can be considered equivalent.

Problem (6.7) is an integer linear program except for constraint (6.7d), which is non-linear due to the log-term in the function $\zeta_{\tau^{\mathrm{MIN}}}^{+}(\gamma)$ as defined in (2.9), the fractional SINR-term and the allocation factor A_{km}. In the following, a linear inner approximation of (6.7d)-(6.7e) based on Sec. 3.2.3 is proposed. Denote a set of I linear functions

$$u_i(\gamma) = \alpha_i \gamma + \beta_i, i = 1, \ldots, I, \tag{6.8}$$

which satisfy the upper bound property

$$\max_i u_i(\gamma) \geq \zeta_{\tau^{\mathrm{MIN}}}^{+}(\gamma) \quad \forall\, \gamma \geq \gamma^{\mathrm{MIN}}, \tag{6.9}$$

as illustrated in Fig. 3.2. Since $f(\gamma)$ in (2.5) is strictly decreasing, all $u_i(\gamma)$ can be designed such that $\alpha_i \leq 0 \ \forall \ i$. To approximate the load for $\gamma \geq \gamma^{\mathrm{MAX}}$, as depicted in Fig. 3.2, a constant function can be used with $u_I(\gamma) = \beta_I = \tau^{\mathrm{MIN}}$. The issue of designing a suitable set of u_i that keep the maximum absolute approximation error below a selectable threshold ϵ is discussed in Sec. 3.2.3.
Let μ_{km} be an optimization parameter designed to be an upper bound of the load term in Eq. (6.7d), such that

$$\mu_{km} \geq u_i(\gamma) \quad \forall \ i, \gamma \geq \gamma^{\mathrm{MIN}} \tag{6.10}$$

and the corresponding matrix $\boldsymbol{\mu} \in \mathbb{R}_{0+}^{K \times M}$. For the interval $\gamma^{\mathrm{MIN}} \leq \gamma \leq \gamma^{\mathrm{MAX}}$, the log-term contained in the function $\zeta^{+}_{\tau^{\mathrm{MIN}}}(\gamma)$ in the constraint (6.7d) is reformulated as

$$\rho_k = \sum_{m=1}^{M} A_{km} \frac{d_m}{W \eta_{km}^{\mathrm{BW}}} \mu_{km} \tag{6.11}$$

where for (6.8)-(6.10)

$$\mu_{km} \geq \alpha_i \frac{\tilde{p}_k g_{km}}{\sum_{j=1,\ldots,K} (1 - \Omega_{jm}) g_{jm} + \sigma^2} + \beta_i \quad \forall \ i, k, m \tag{6.12}$$

The product of μ_{km} and allocation parameter A_{km} is in the following as $\Lambda_{km} = \mu_{km} A_{km}$ and the corresponding matrix as $\boldsymbol{\Lambda} \in \mathbb{R}_{0+}^{K \times M}$. This bilinear product formulation for $\boldsymbol{\Lambda}$ is replaced by a linear reformulation using (3.4) by adding the constraint that $(\mu_{km}, \beta_1, A_{km}, \Lambda_{km}) \in \mathcal{L}$.
In order to approximate the interference levels in the denominator of the SINR term Eq. (2.2), the fractional bounding discretization outlined in Sec. 3.2.4. Scalar interference levels Ψ_{nkm} are introduced with interference scenario index $n = 1, \ldots, N$, and the corresponding three-dimensional scalar tensor $\boldsymbol{\Psi} \in \mathbb{R}_{0+}^{N \times K \times M}$. Denote a binary interference scenario selection parameter ϕ_{nkm} and the corresponding three-dimensional binary tensor $\boldsymbol{\phi} \in \{0,1\}^{N \times K \times M}$. To ensure that the solution of the approximate problem is always feasible for the original, a constraint is added that the selected discrete interference level is always an over-approximation of the actual interference:

$$\sum_{n=1}^{N} \phi_{nkm} \Psi_{nkm} \geq \sum_{j=1,\ldots,K} (1 - \Omega_{jm}) g_{jm} + \sigma^2 \ \forall k, m \tag{6.13}$$

When implementing the selection parameter $\boldsymbol{\phi}$ in Eq. (6.12), the bilinear product $\tilde{p}_k g_{km} \phi_{nkm}$ is replaced with an auxiliary parameter, for which the lifting variable $\Phi_{nkm} = \tilde{p}_k g_{km} \phi_{nkm}$ is introduced with the corresponding tensor variable $\boldsymbol{\Phi} \in \mathbb{R}_{0+}^{N \times K \times M}$. Again, the product computation of $\boldsymbol{\Phi}$ will be replaced by an auxiliary parameter using (3.4) by adding the constraint $\left(\tilde{p}_k g_{km}, P_k^{\mathrm{MAX}} g_{km}, \phi_{nkm}, \Phi_{nkm}\right) \in \mathcal{L} \ \forall n, k, m$.

The proposed linear inner approximation of (6.7) is the following:

$$\underset{\boldsymbol{z},\tilde{\boldsymbol{p}},\boldsymbol{A},\tilde{\boldsymbol{\rho}},\boldsymbol{\Omega},\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{\phi},\boldsymbol{\Phi}}{\text{minimize}} \quad \sum_{k=1}^{K} \Gamma\left(z_k, \tilde{p}_k, \tilde{\rho}_k\right) \tag{6.14a}$$

$$\text{subject to} \quad (6.5c) - (6.5d), (6.6b), (3.4), (6.7b) - (6.7c), (6.7f), (6.13)$$

$$\tilde{\rho}_k = \sum_{m=1}^{M} \left(\frac{d_m}{W \eta^{\mathrm{BW}}} \Lambda_{km} \right) \ \forall\, k \tag{6.14b}$$

$$\tilde{\rho}_k \leq 1 \ \forall\, k \tag{6.14c}$$

$$\sum_{n=1}^{N} \phi_{nkm} = 1 \ \forall\, k, m \tag{6.14d}$$

$$\mu_{km} \geq \alpha_i \sum_{n=1}^{N} \frac{\Phi_{nkm}}{\Psi_{nkm}} + \beta_i \ \forall\, i, k, m \tag{6.14e}$$

$$(\mu_{km}, \beta_1, A_{km}, \Lambda_{km}) \in \mathcal{L} \ \forall\, k, m \tag{6.14f}$$

$$\left(\tilde{p}_k g_{km}, P_k^{\mathrm{MAX}} g_{km}, \phi_{nkm}, \Phi_{nkm}\right) \in \mathcal{L} \ \forall\, n, k, m \tag{6.14g}$$

$$z_k, A_{km}, \phi_{nkm} \in \{0, 1\} \ \forall\, n, k, m \tag{6.14h}$$

$$\tilde{p}_k, \Omega_{km}, \mu_{km}, \Lambda_{km}, \Phi_{nkm} \in \mathbb{R}_{0+} \ \forall\, n, k, m \tag{6.14i}$$

Proposition 6.3.1. *Problem* (6.14) *is an inner approximation of problem* (6.7)*, i.e. for every point* $\{\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{A}\}$ *solving* (6.14) *a feasible point of* (6.7) *can be constructed.*

Proof. The transmit power constraints (6.6b), the allocation constraints (6.5c)-(6.5d) and the signal power constraints (6.7b)-(6.7c) are identical in problem (6.7) and (6.14). The proposition therefore holds if the load in (6.14b) is an inner approximation of that in (6.7d), specifically if

$$\sum_{m=1}^{M} \left(\frac{d_m}{W \eta^{\mathrm{BW}}} \Lambda_{km} \right) \geq \sum_{m=1}^{M} A_{km} \frac{d_m}{W \eta_{km}^{\mathrm{BW}}} \frac{1}{\log_2\left(1 + \frac{\tilde{p}_k g_{km}}{\sum_{j=1,\ldots,K} (1 - \Omega_{jm}) g_{jm} + \sigma^2}\right)} \quad \forall\, k. \tag{6.15}$$

Due to (6.14f), there is $\Lambda_{km} = \mu_{km} A_{km}$, therefore (6.15) is satisfied if

$$\mu_{km} \geq \frac{1}{\log_2\left(1 + \frac{\tilde{p}_k g_{km}}{\sum_{j=1,\ldots,K} (1 - \Omega_{jm}) g_{jm} + \sigma^2}\right)} \quad \forall\, k, m, \tag{6.16}$$

from which, with (6.14e) and (6.9) applied to the left- and right-hand side of Eq. (6.16),

respectively,

$$\alpha_i \sum_{n=1}^{N} \frac{\Phi_{nkm}}{\Psi_{nkm}} + \beta_i \geq \alpha_i \sum_{n=1}^{N} \frac{\tilde{p}_k g_{km}}{\sum_{j=1,\dots,K}(1-\Omega_{jm})g_{jm} + \sigma^2} + \beta_i \quad \forall\, i,k,m. \tag{6.17}$$

Due to the constraints (6.14g), which implement the bilinear constraint $\Phi_{nkm} = \tilde{p}_k g_{km} \phi_{nkm}$, and due to $\phi_{nkm} \in \{0,1\}\ \forall\, n,k,m$, there is

$$\sum_{n=1}^{N} \frac{\Phi_{nkm}}{\Psi_{nkm}} = \frac{\tilde{p}_k g_{km}}{\sum_{n=1}^{N} \phi_{nkm} \Psi_{nkm}} \quad \forall\, n,k,m. \tag{6.18}$$

Substituting (6.18) in the left-hand side of (6.17) yields the inequality

$$\alpha_i \sum_{n=1}^{N} \frac{\tilde{p}_k g_{km}}{\sum_{n=1}^{N} \Psi_{nkm}} + \beta_i \geq \alpha_i \sum_{n=1}^{N} \frac{\tilde{p}_k g_{km}}{\sum_{j=1,\dots,K}(1-\Omega_{jm})g_{jm} + \sigma^2} + \beta_i \quad \forall\, i,k,m, \tag{6.19}$$

which holds due to the constraint (6.13) for $\alpha_i \leq 0\ \forall\, i$, thus proving the proposition. □

The tightness of the approximating problem (6.14) with regards to problem (6.7) depends on two factors. The first factor is related to how closely the linear functions u_i approximate the load function as in Eq. (6.9). The second factor is how closely the discrete interference levels Ψ_{nkm} approximate the actual interference level $\sum_{j=1,\dots,K}(1-\Omega_{jm})g_{jm} + \sigma^2$. Proposition 6.3.1 holds irrespectively the choice of the discrete interference levels Ψ_{nkm}. Certain changes in interference levels, specifically the removal of strongest interferers, cause large differences in the load caused by a DP. The levels Ψ_{nkm} can be chosen in such a way that these changes can be reflected by the selection of a different interference scenario. The accuracy of the inner approximation can be improved by using a larger number of interference levels, at the cost of increased problem complexity.

Based on the considerations for the fractional bounding discretization approach discussed in Sec. 3.2.4, introduce for each pair (m,k) of DP m allocated to cell k, interference levels Ψ_{nkm} that mainly reflect transmit power changes of the first- and second-strongest interferers [MHV+12, RCBHP17b, GKN+15]. With the indices defined in Eqs. (2.19) and (2.20), the interference levels are computed as

$$\Psi_{nkm} = L_n^{\mathrm{P}} p_{\kappa_{km}^{\mathrm{P}}} g_{\kappa_{km}^{\mathrm{P}} m} + L_n^{\mathrm{S}} p_{\kappa_{km}^{\mathrm{S}}} g_{\kappa_{km}^{\mathrm{S}} m} + L_n^{\mathrm{R}} \sum_{j \setminus \{\kappa_{km}^{\mathrm{P}}, \kappa_{km}^{\mathrm{S}}, w\}} p_j g_{jm} + \sigma^2, \tag{6.20}$$

where the parameters L_n^{P},L_n^{S} and L_n^{R} denote the weighting factors for primary-, secondary- and remaining interferers, respectively. Keeping in mind that the focus

Table 6.1. Weighting factors for computation of interference scenarios Ψ_{nkm}, used for an over-approximation of the actual interference level.

$n =$	1	2	3	4	5	6	7
L_n^{P}	1	0.75	0.5	0.25	0	0	0
L_n^{S}	1	1	1	1	1	0	0
L_n^{R}	1	1	1	1	1	1	0

is on transmit power changes for the first- and second strongest interferers, a suitable set of weighting factors to compute the interference levels Ψ_{nkm} is shown, for example, in Table 6.1.

6.4 Simulation Results

To evaluate the performance of the proposed method, a heterogeneous wireless communication network is simulated containing 4 macro- and 4 pico cells as illustrated in Fig. 6.1. The selected system parameters are summarized in Table 6.2. The selectable transmit power range and antenna gains are chosen as 36dBm − 46dBm with 15dB antenna gain for macro cells and 26dBm − 36dBm with 5dB antenna gain for small cells. A bias value of $\theta_k = 3$dB is used for small cells to slightly increase their coverage area. The proposed method using Problem (6.14) was solved using CVX for MATLAB [GB14, GB08] and Gurobi as a MILP solver [GUR].

6.4.1 Energy Consumption Modeling Comparison

To evaluate the effect of different models for the cell energy consumption Eq. (6.4), four scenarios with different weighting factors are introduced in Sec. 6.4.1. Equal weighting for all energy consumption components is used in Model 1. For Model 2,3 and 4 the load-, transmit power- and on-off-status-based components are respectively with zero, and use equal weighting between the remaining two components.

First the effect of different parameter settings is evaluated for the network illustrated in Fig. 6.1, but without the four pico-cells. A comparison of the energy consumptions for the described models is shown in Fig. 6.2. It is observable that Model 4, where

Table 6.2. Simulation parameters of a downlink LTE network for energy consumption minimization. The transmit power of the cells is optimized inside a 10dB interval. Results are averaged over 5000 simulations with fixed base station positions and randomly distributed DPs.

Area size	1000×1000 m
Noise power	-145 dBm/Hz
System bandwidth W	20 MHz
Position of macro BS	MBS1 at [200m, 200m] MBS2 at [150m, 850m] MBS3 at [800m, 230m] MBS4 at [780m, 820m]
MBS transmit power range $P^{\mathrm{MIN}} \dots P^{\mathrm{MAX}}$	36dBm $\dots$ 46dBm
MBS antenna gain $\tilde{g}^{\mathrm{ABS}}$	15dB
MBS bias value θ_k	0dB
Position of pico BS	PBS1 at [500m, 700m] PBS2 at [520m, 310m] PBS3 at [320m, 500m] PBS4 at [690m, 490m]
PBS transmit power range $P^{\mathrm{MIN}} \dots P^{\mathrm{MAX}}$	26dBm $\dots$ 36dBm
PBS antenna gain $\tilde{g}^{\mathrm{ABS}}$	5dB
PBS bias value θ_k	3dB
DP antenna gain $\tilde{g}^{\mathrm{ADP}}$	0dB
Propagation loss $\tilde{g}^{\mathrm{PATH}}$	3GPP TS 36.814 [3GP16]
Bandwidth efficiency η^{BW}	0.8
SINR requirement γ^{MIN}	-10dB
SINR threshold γ^{MAX}	20dB

Table 6.3. Weighting factors for different models of $\Gamma\left(x_k, \tilde{p}_k, \rho_k\right)$

Model nr.	1	2	3	4
κ_1	0.33	0.5	0.5	0
κ_2	0.33	0.5	0	0.5
κ_3	0.33	0	0.5	0.5

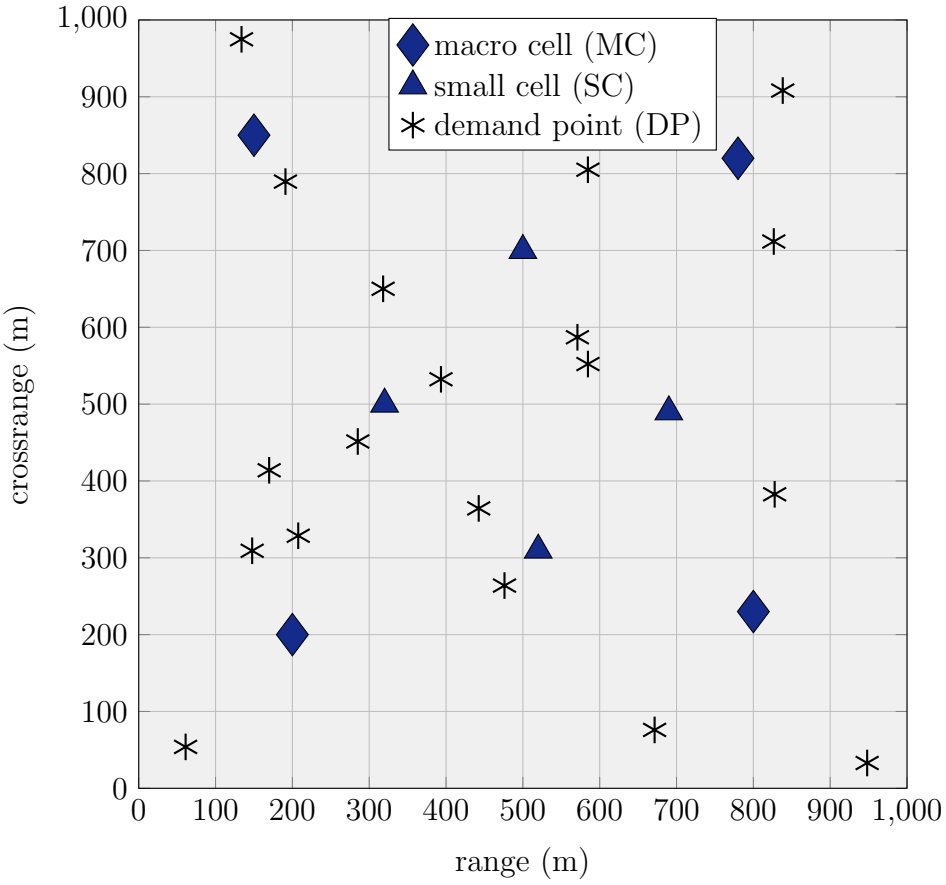

Figure 6.1. Illustration of the network scenario for energy consumption optimization with 4 macro- and 4 small cells and an example distribution of 20 DPs.

the energy consumption solely depends on the variable parameters transmit power and cell load, has lower energy consumption. This is rather plausible, as there is no option for the algorithm to decrease the fixed energy consumption of active cells in Models 1-3. For these models, the energy consumption increases of course with the data demand. But between the models the energy consumption also increases with higher weighting factors κ_3 for the load-based energy consumption and with lower κ_2 for the transmit-power-based energy consumption. This indicates that there is more flexibility for the algorithm to decrease transmit power than to decrease cell load. Also, the energy consumption in Models 1-3 is rather similar for demands up to about 2 Mbit/s, which identifies the range where the network can mostly be operated with only one active cell. This is confirmed in Fig. 6.3, where the number of active cells is shown for each model. Model 4 uses on average more active cells than Models 1-3, because it does not consider a fixed power consumption for active cells. Models 1-3 use approximately the same number of active cells.

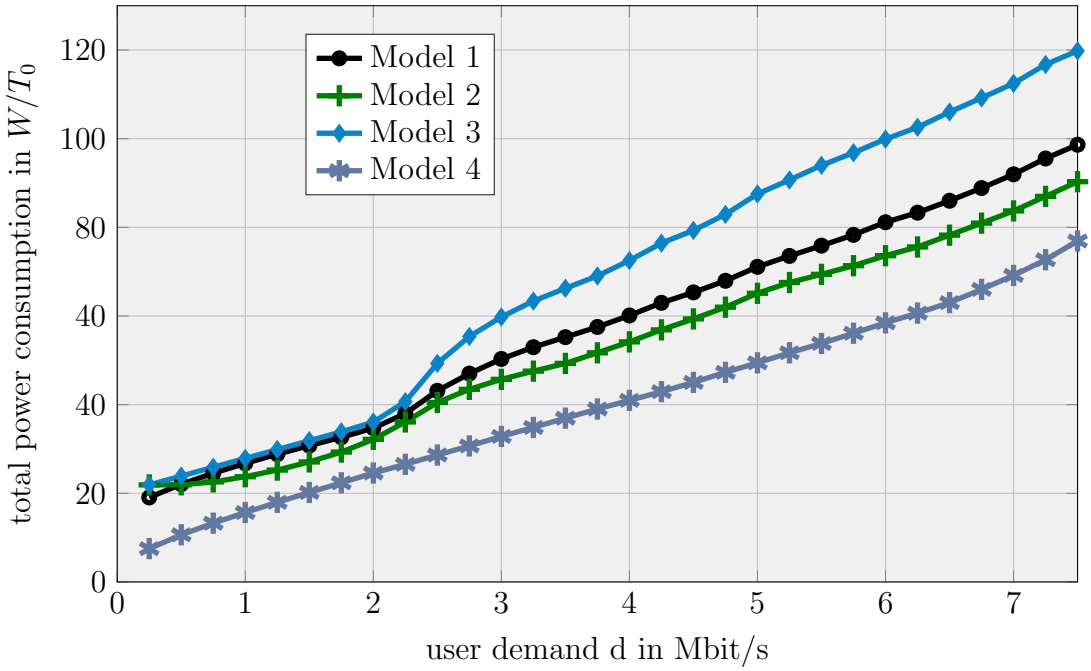

Figure 6.2. Energy consumption for different energy consumption models, 4 macro cells, M=20 DPs, averages of 250 simulations. Only Model 4, where the cell activity status does not contribute to the energy consumption, uses on average a higher number of cells.

6.4.2 Performance Comparison of Schemes

In the following, for the energy consumption modeling of cells, Eq. (6.4) is used with $\kappa_1 = 0.5$, $\kappa_2 = 0.5$ and $\kappa_3 = 0$. This implies that the power consumption of cell k depends on its on-off status indicator z_k and its transmit power p_k. The power consumption is modeled this way in order to allow comparability of the proposed MILP with an established heuristic method proposed in [HYLS15] that focuses on transmit power minimization. As a performance benchmark for our energy minimization algorithm the power scaling method introduced in [HYLS15] is used, which is extended in the following ways to make it applicable to our problem: power scaling is used for all possible configurations of all cells' on-off status $\boldsymbol{z}$. Resulting transmit powers obtained by the algorithm of [HYLS15] that lie below or above the bounds specified in Table 6.2 are projected to the lower- and upper bound respectively. Then, the best configuration that does not violate load constraints is selected as the solution. This

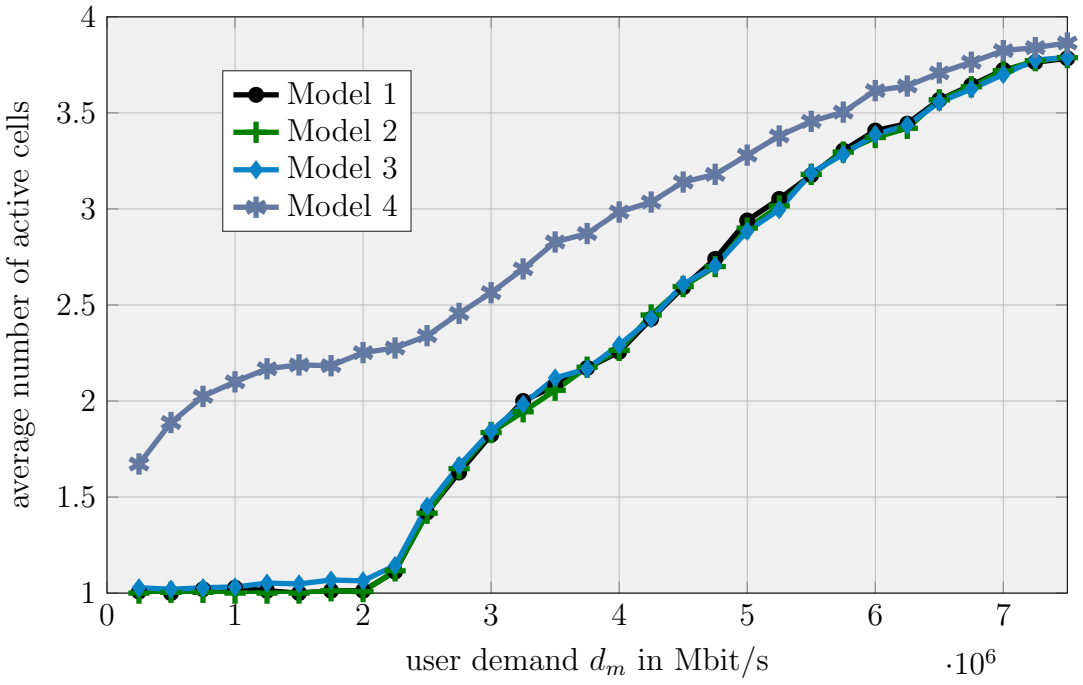

Figure 6.3. Number of active cells for different energy consumption models, 4 macro cells, M=20 users, averages of 250 simulations. The achieved Energy consumption follows similar patterns for all models.

algorithm therefore combines an exhaustive search over all configurations for $\boldsymbol{z}$ with power scaling being used in each configuration. It is in the following in all figures denoted as "power scaling + exh. search". The second approach used for comparison is an exhaustive search over all combinations of cells being switched on or off, with the transmit powers being fixed to P^{MAX}, which in the following is indicated as "max power cell switching". The solution of the original MINLP in (6.5) is unsuitable as a lower bound solution even for small problem sizes, because even for fixed binary optimization parameters the resulting continuous problem is still nonconvex. Deploying $M = 20$ DPs randomly in the network area illustrated in Fig. 6.1, 5000 network scenarios are generated and each DPs data demand in each scenario is scaled between $d_m = 0.25$Mbit/s and $d_m = 7.5$Mbit/s. The proposed energy-minimized solution obtained from solving problem (6.14) is compared to the solutions of the aforementioned max. power cell switching and combined power scaling and exhaustive search methods [HYLS15]. The probability of obtaining a feasible solution with no overloaded cells is illustrated in Fig.6.4. The proposed MILP based method is much more likely to find a feasible and power-minimized solution even in high demand scenarios.

Out of the 5000 evaluated scenarios, only those can be considered where the original

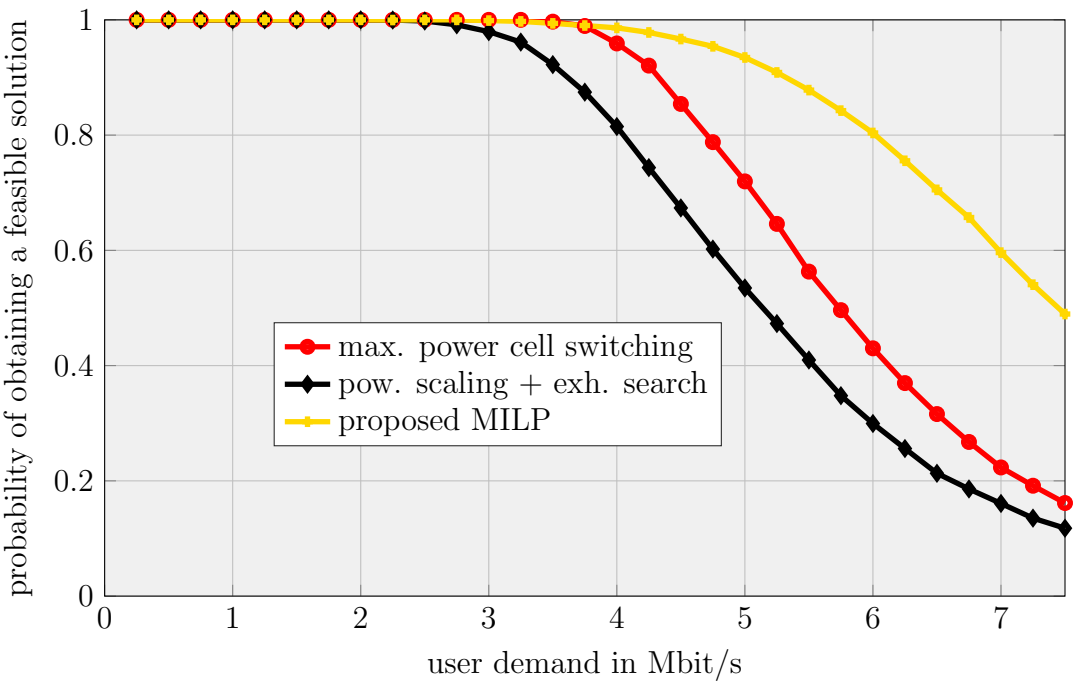

Figure 6.4. Probability of obtaining a feasible solution over increasing user demand, evaluated over 5000 simulations of $M = 20$ randomly distributed demand points. The proposed MILP-based scheme achieves the highest solution percentage.

configuration with maximum transmit power is a feasible solution to the energy minimization problem. In this case, the transmission at full power can be considered as a fallback solution for each energy minimization scheme, should it fail to find an "energy-minimized solution". In this case, for each scheme an average between 5000 datapoints can be computed. The achieved average energy consumption for each method is shown in Fig. 6.5. It is observable that the proposed method greatly benefits from its increased chances of finding a feasible solution. In the following, it has to be determined if this superior performance is still present if the effect of the solving percentage is not present.

To ensure a fair comparison, the respective averages of performance indicators will in the following be computed only from those scenarios that were solved by all methods. The following performance indicators are discussed: energy consumption, cell load, and number of active cells. Fig. 6.6 shows the average power consumption achieved by each of the three considered energy minimization schemes. The proposed MILP-based approach achieves lower power consumption levels than both the cell switching and the heuristic approach. The cell switching method noticeably achieves good performance up until about 3Mbit/s, with the performance significantly deteriorating for higher

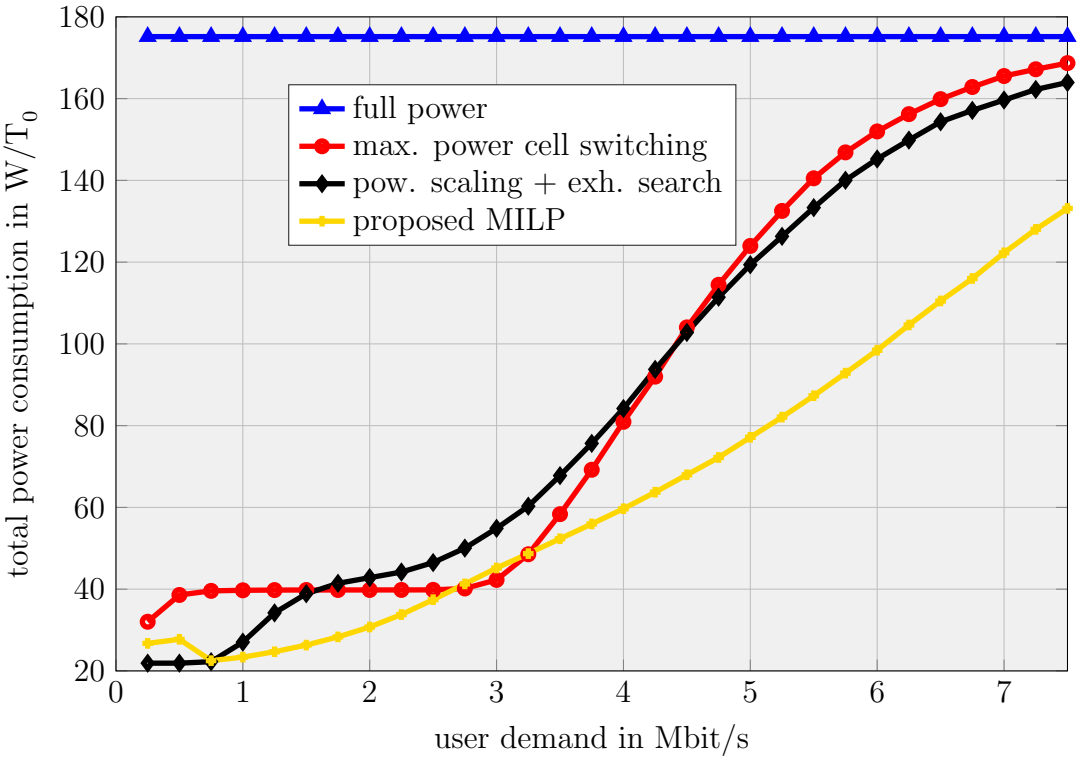

Figure 6.5. Energy consumption for energy minimization schemes over increasing user demand, averaged over 5000 simulations of $M = 20$ randomly distributed demand points. Each method uses maximum transmit power as a fallback solution The proposed scheme achieves the lowest average energy consumption levels of the evaluated schemes.

demands.

In Fig. 6.7, the average number of active cells is shown. For very low demands, it can be observed that the number of cells is not increasing continuously with the demand, as the proposed algorithm for some scenarios serves all users exclusively with pico cells, instead of using a single macro cell. In practice this does not pose a problem since for these low load levels offloading is not required. On average however less than 4 cells are being used, showing that small cells are only used sporadically or for low demand levels. For very high demand levels, the proposed method utilizes the lowest number of cells.

The average load factor of active cells is shown in Fig. 6.8. It is observable that the cell load does not converge to 1 even for high loads. It was shown in [HYLS15] that for minimum energy consumption, the load would be equal to 1. This however only holds if the transmit power can be increased or decreased without bounds (i.e. for $P^{\mathrm{MIN}} = 0$

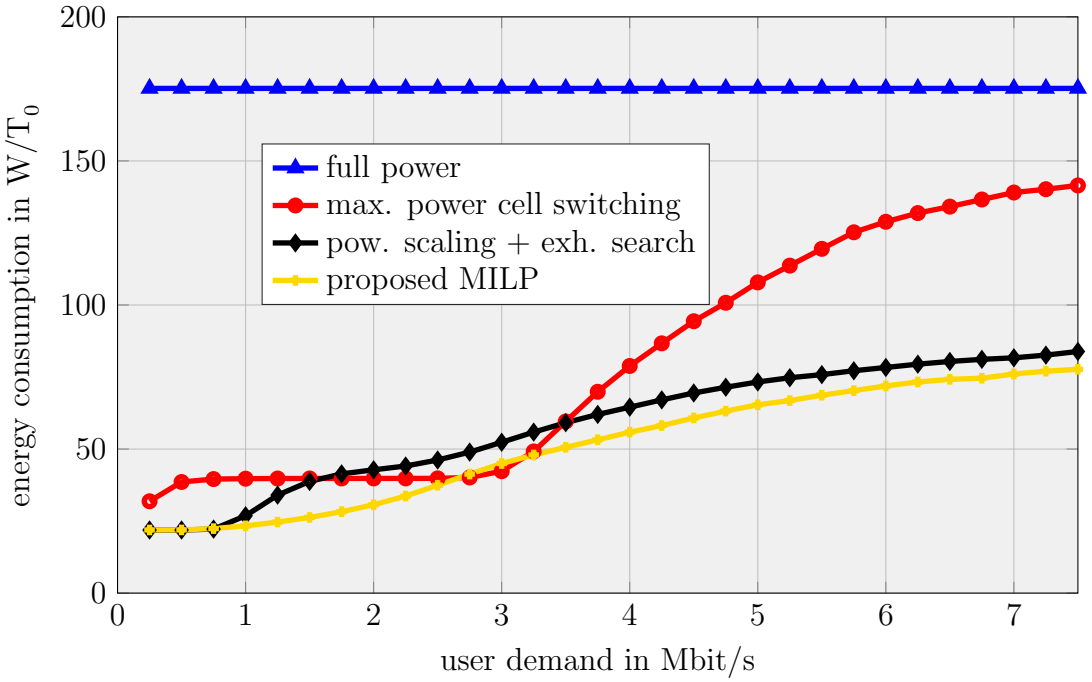

Figure 6.6. Energy consumption for energy minimization schemes over increasing user demand, averaged over 5000 simulations of $M = 20$ randomly distributed demand points. The proposed scheme achieves the lowest average energy consumption levels of the evaluated schemes.

and $P^{\mathrm{MIN}} = \infty$), and if the cell load is a strictly decreasing function of the transmit power. With the upper- and lower bounds on the transmit power, the discontinuities introduced in the load computation, and the user allocation changing dynamically with the transmit powers,it can be observed from Fig. 6.8 that this property no longer holds. The energy consumption is also evaluated for a varying number of users as shown in Fig. 6.9. The network is simulated in 200 scenarios with each between 5 and 25 users with a demand of 6 Mbit/s each. Only scenarios are used that were solved by all methods, and it shows that the results are qualitatively similar to the simulation presented in Fig. 6.6. This has the strong implication that the energy saving capabilities of the proposed methods depend on the density of data demand per area, but not on the actual number of DPs.

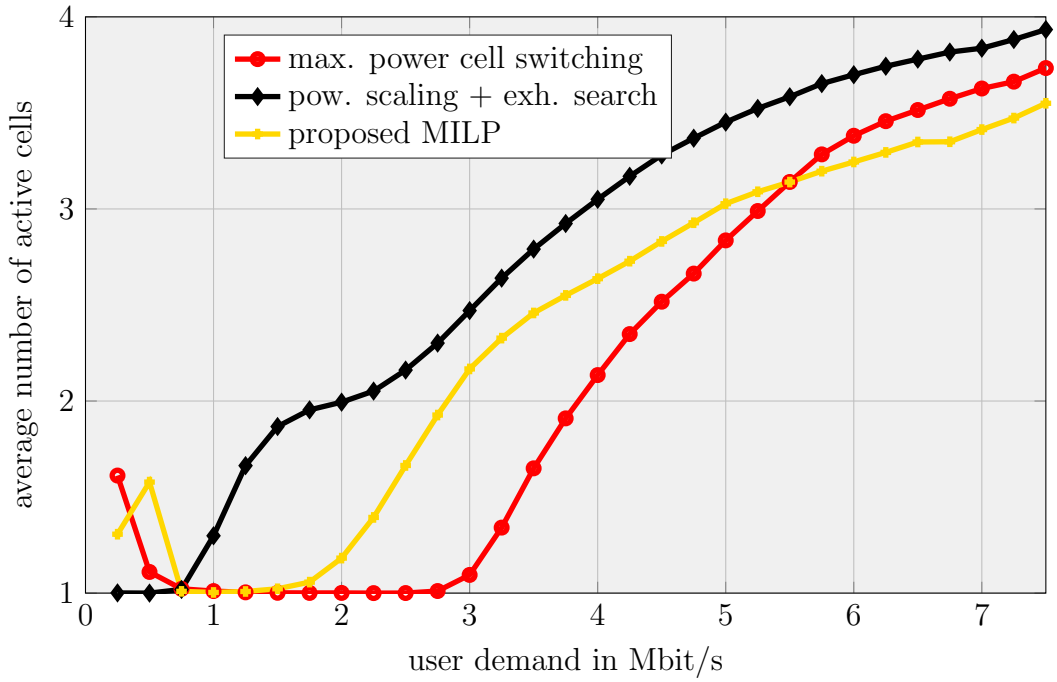

Figure 6.7. Number of active cells for energy minimization schemes over increasing user demand, averaged over 5000 simulations of $M = 20$ randomly distributed demand points. For high demand, the proposed scheme on average utilizes the lowest number of cells.

6.5 Summary

This chapter addressed the challenge of minimizing the energy consumption of a wireless communication network by joint optimization of the base station transmit power and the cell activity. A mixed-integer nonlinear optimization problem is formulated, for which a computationally tractable linear inner approximation algorithm was provided. The proposed method offers great flexibility in optimizing the network operation by considering multiple system parameters jointly, which mitigates a major drawback of existing state-of-the-art schemes that are mostly based on heuristics. Multiple simplifications used in other state of the art methods to allow the application of heuristic schemes are not required in the proposed method.

The simulation results show that the proposed approach achieved a further decrease in energy consumption relative to both an optimization of the cell activity and a comparable heuristic method. Additionally, it achieves a higher success rate in finding an operable solution for high-demand network scenarios.

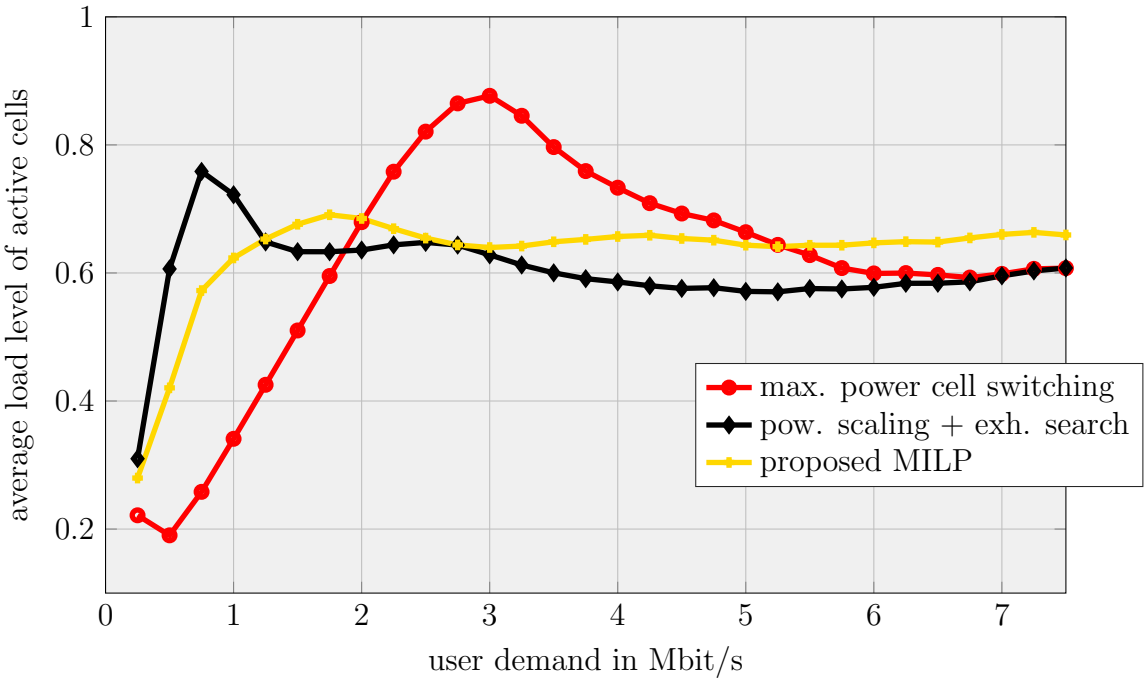

Figure 6.8. Load of active cells for energy minimization schemes over increasing user demand, averaged over 5000 simulations of $M = 20$ randomly distributed demand points.

Even though the proposed method consists in linear approximations of the originally mixed integer nonlinear program with bilinear and nonconvex constraints, it still yields very high complexity, making it impractical for the optimization of large networks. Further work could be dedicated to combining existing heuristic methods with an utilization of the proposed approach to optimize smaller clusters of the network, to allow for better scalability.

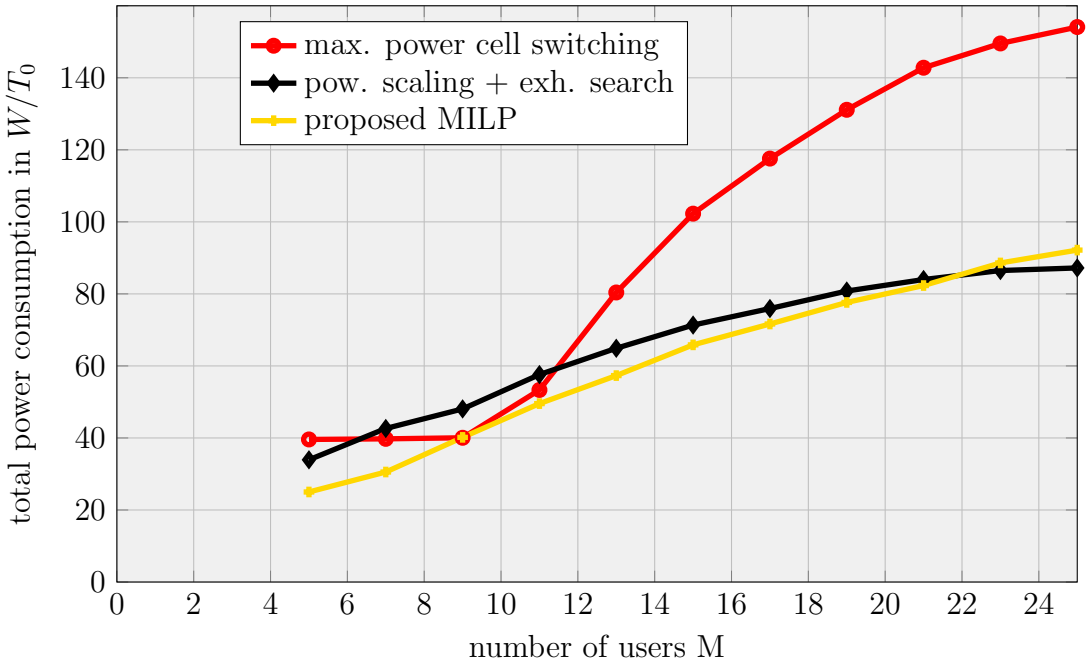

Figure 6.9. Energy consumption for energy minimization schemes over increasing number of DPs M, averaged over 200 simulations with a data demand d_m of 6 Mbit/s for each DP. The similarity to increasing DP demand implies that the algorithm performance depends on the spatial demand density.

Chapter 7

Decentralized Load Balancing

7.1 Introduction and Contributions

In addition to the correct placement and activity scheduling of SCs discussed in Chapter 4, the optimized allocation of users to MCs or SCs is a subject of current research [YY17, SY12b, YRC+13, AWFT15]. The allocation can be optimized while the network is in operation, or allocation rules can be devised before, based on demand forecasts. Most prominently, small cell range expansion has been proposed as an effective way to move users from the typically overloaded MCs to the less utilized SCS. This can be achieved either by optimizing the user allocation directly (for example by solving (2.13)) or by introducing the cell bias values discussed in Sec. 2.2 to the signal power report received by the user node. For the allocation decision with range expansion, the signal powers from small cells are increased with a bias value. This leads to more users being allocated to SCs, which corresponds to an increased coverage area. The main parameter to be optimized for range expansion is the bias value for each SC.

In this chapter, approaches for in-operation load balancing based on allocation optimization and range expansion optimization are both solved using decentralized learning-based approaches. These learning-based approaches are used to mitigate the aforementioned problem of high communication and coordination overhead with established methods. The proposed methods rely on utilizing the multi-class SVMs discussed in Sec. 3.3 as resource-allocation tools. For the training of these SVMs, local attributes of the network state are extracted that are easily accessible to each DP and cell, such as current load levels, channel conditions and the neighboring cell topology. The small cell attributes and the optimal MILP results are used to train a classifier based on multi-class support vector machines. This classifier is then applied locally in each DP to allocate to a suitable neighboring cell or in each SC to find its optimal bias value in new network scenarios. Using machine learning classifiers as improvised resource allocation schemes in wireless communication networks is only being considered recently, and comparable methods have not yet been introduced.

7.1.1 State-of-the-Art

The problem of allocating DPs to cells such that load balancing between the cells is achieved is formulated in Eqs. (2.13). Other allocation schemes have been proposed [YY17, AWFT15]. The authors in [YY17] propose a load balancing problem similar to Eqs. (2.13), but use the cell load as a weighting factor for the interference that each respective cell creates. This complicates the optimization drastically, but as discussed in 2.3, is not a worst-case assumption and therefore this weighting factor is not used in the system model in this thesis. In [AWFT15], a load balancing scheme is proposed based on a ILP that has the same structure as Eqs. (2.13). Optimized schemes for range expansion have been proposed in [SY12b, YRC+13]. The problem formulation in [SY12b] again uses the aforementioned weighting factor, and utilizes piecewise linearization to obtain optimized bias values for range expansion. The authors in [YRC+13] propose to solve the user association problem such that a utility function based on the achieved rates is maximized. While the underlying user association problem for load balancing is a (combinatorial) IP, the common drawback of the established optimization schemes is that they require extensive knowledge about the channel conditions and the state of each network entity to perform the bias value and allocation optimization, which is carried out either centrally or using consensus algorithms.

7.1.2 Contributions and Overview

The key contributions in this chapter can be summarized as follows: Multi-class SVMs are utilized as tools to obtain decentralized solutions for ILPs characterizing network load balancing problems. The training of the required SVM-based multiclass is performed using historical network data, but once training is complete, the resulting classifiers can be applied in a fast and decentralized scheme during network operation. Both the proposed decentralized DP-allocation scheme and range expansion scheme achieve close to (globally) optimal performance.

The remainder of this chapter is organized as follows: The proposed methods for load balancing based on DP allocation and range expansion are explained in Secs. 7.2 and 7.3, respectively. In Sec. 7.4 simulation results are provided, followed by a final summary and conclusion in Sec. 7.5.

7.2 User Allocation Optimization

The primary, secondary and tertiary allocation candidates of DP m are, in descending order of magnitude, those cells which can provide the first-, second-, and third-highest signal power $p_k g_{km}$ at the DP's location. Their indices are listed in the vectors $\boldsymbol{\kappa}^{\mathrm{P}}, \boldsymbol{\kappa}^{\mathrm{S}}, \boldsymbol{\kappa}^{\mathrm{T}} \in \mathbb{N}^{M\times 1}$, respectively, with their respective elements determined as

$$\kappa_m^{\mathrm{P}} = \arg\min_{k} p_k g_{km}, \tag{7.1}$$

$$\kappa_m^{\mathrm{S}} = \arg\min_{k\setminus\{\kappa_m^{\mathrm{P}}\}} p_k g_{km}, \tag{7.2}$$

$$\kappa_m^{\mathrm{T}} = \arg\min_{k\setminus\{\kappa_m^{\mathrm{P}},\kappa_m^{\mathrm{S}}\}} p_k g_{km}. \tag{7.3}$$

As shown for Lemma 2.3.1, in the case that $A_{km} = 1$ for $k = \kappa_m^{\mathrm{P}}$ for all DPs m, each DP is allocated according to the cell providing the connection with the highest SINR. In this way the additional load caused by each individual connection is minimized. This common allocation scheme is used as a heuristic baseline approach to evaluate the subsequent load balancing solutions, in the following denoted as "max. SINR".

In the following, assume that $\boldsymbol{A}^*$ represents the DP-cell allocation solution obtained from solving problem (2.13). Denote the label vector $\boldsymbol{y} \in \mathbb{N}^{M\times 1}$ which has elements y_m determined as follows:

$$y_m = \begin{cases} 2 \text{ if } A^*_{\kappa_m^{\mathrm{S}} m} = 1 \\ 3 \text{ if } A^*_{\kappa_m^{\mathrm{T}} m} = 1 \\ 1 \text{ otherwise} \end{cases} \tag{7.4}$$

For the training of the proposed statistical learning approach for user allocation, attributes need to be extracted for the three candidate allocation cells of each user m. These attributes are designed to reflect specific knowledge of the network and the parameters deemed significant for the allocation problem. Three attributes are extracted for each cell. The first attribute is an indicator of cell type defined as

$$F^{\mathrm{TYPE}}(k) = \begin{cases} 1 & \text{if cell } k \text{ is a small cell,} \\ 0 & \text{otherwise.} \end{cases} \tag{7.5}$$

The second attribute describes the additional load that user m would cause to cell k if it was allocated to it, based on the load definitions introduced in Sec. 2.2:

$$F^{\mathrm{LOAD}} = \frac{d_m}{R_{km}} \tag{7.6}$$

The third attribute is the "would-be" load of cell k, if the max. SINR scheme was being used. This attribute serves as a measure of load caused by DPs in each cell's coverage area. The third attribute is determined as follows:

$$F^{\mathrm{COV}}(k) = \sum_{m|\kappa_m^{\mathrm{P}}=k} F^{\mathrm{LOAD}}(k,m). \tag{7.7}$$

Using all three of the aforementioned attributes for each of the three candidate cells for allocation, the attribute vector of DP m is determined as

$$\begin{aligned}\boldsymbol{h}_m = \big[&F^{\mathrm{TYPE}}(\kappa_m^{\mathrm{P}}), F^{\mathrm{TYPE}}(\kappa_m^{\mathrm{S}}), F^{\mathrm{TYPE}}(\kappa_m^{\mathrm{T}}), F^{\mathrm{LOAD}}(\kappa_m^{\mathrm{P}},m), F^{\mathrm{LOAD}}(\kappa_m^{\mathrm{S}},m), \dots \\ &F^{\mathrm{LOAD}}(\kappa_m^{\mathrm{T}},m), F^{\mathrm{COV}}(\kappa_m^{\mathrm{P}}), F^{\mathrm{COV}}(\kappa_m^{\mathrm{S}}), F^{\mathrm{COV}}(\kappa_m^{\mathrm{T}})\big]^\top. \end{aligned} \tag{7.8}$$

This results in a training problem of a multi-class classifier which will be solved by training two SVMs, with the first being used to identify DPs that are allocated to their secondary cell, and the second SVM identifying those that are allocated to their tertiary cell, characterized by the normal vectors to their separating hyperplanes $\boldsymbol{\omega}^{21}$ and $\boldsymbol{\omega}^{31}$ respectively. The used soft-margin SVM is discussed in Sec. 3.3.2. Based on the one-versus-one multiclass extension to the SVM discussed in Sec. 3.3.3, denote as $\hat{y}_m$ the cell type that is classified by the SVMs according to the two decision functions based on a feature observation $\hat{\boldsymbol{h}}$, which is computed as:

$$\hat{y}_m = \begin{cases} 2 \text{ if } (\boldsymbol{\omega}^{21})^\top \vartheta(\hat{\boldsymbol{h}}_m) + b^{21} \geq 0 \text{ and} \\ \quad (\boldsymbol{w}^{21})^\top \vartheta(\hat{\boldsymbol{h}}_m) + b^{21} \geq (\boldsymbol{w}^{31})^\top \vartheta(\hat{\boldsymbol{h}}_m) + b^{31} \\ 3 \text{ if } (\boldsymbol{w}^{31})^\top \vartheta(\hat{\boldsymbol{h}}_m) + b^{31} \geq 0 \text{ and} \\ \quad (\boldsymbol{w}^{31})^\top \vartheta(\hat{\boldsymbol{h}}_m) + b^{31} \geq (\boldsymbol{w}^{21})^\top \vartheta(\hat{\boldsymbol{h}}_m) + b^{21} \\ 1 \text{ otherwise} \end{cases} \tag{7.9}$$

Using Eq. (7.9), the allocation decisions for all DPs m in a given network scenario is determined, which leads to a load-balanced allocation solution for the full network.

7.3 Allocation Bias Optimization

For the optimization of cell range expansion, a set of available nonnegative bias values is denoted as $\mathcal{S} = \delta_1, \dots, \delta_{\tilde{S}}$. The K-element vector of the selected bias values for all cells is denoted as $\boldsymbol{\theta}$ with the elements $\theta_k \in \mathcal{S}_k$. For example, if SCs operate with any of the available bias values and MCs would not be biased, there would be $\theta_k = 1\ \forall k \in \mathcal{C}^{\mathrm{MC}}$ and $\theta_k \in \mathcal{S}\ \forall k \in \mathcal{C}^{\mathrm{SC}}$ In the following, denote the $\boldsymbol{A}$ obtained from applying (2.14)

with $\theta_k = 1\ \forall k$ (no bias) as $\tilde{\boldsymbol{A}}$. Similarly, $\boldsymbol{A}^0$ denotes the allocation result according to Eq. (2.14) for $\theta_k = 1\ \forall k \in \mathcal{C}^{\mathrm{MC}}$ (no bias) and $\theta_k = 0\ \forall k \in \mathcal{C}^{\mathrm{SC}}$ (bias).

In the following a scheme is introduced to find the optimal bias values for each cell that minimize the maximum load of any cell in the network. These optimal bias values are obtained as the optimal solution of a mixed integer problem. Assume that $\theta_k \in \mathcal{S}_k\ \forall k \in \mathcal{C}^{\mathrm{SC}}$ and $\theta_k = 1\ \forall k \in \mathcal{C}^{\mathrm{MC}}$, which means that SCs can operate with any of the available bias values and MCs operate without bias. The proposed problem can be formulated as follows:

$$\underset{\Pi, \boldsymbol{A}, \boldsymbol{\theta}}{\text{minimize}} \quad \Pi \tag{7.10a}$$

$$\text{subject to} \quad \Pi \geq \sum_{m=1}^{M} A_{km} \frac{d_m}{R_{km}}\ \forall k \tag{7.10b}$$

$$\sum_{k} A_{km}\theta_k p_k g_{km} \geq (1 - A_{jm})\theta_j p_j g_{jm}\ \forall j, m \tag{7.10c}$$

$$\sum_{k=1}^{K} A_{km} = 1\ \forall m \tag{7.10d}$$

$$\Pi \in \mathbb{R}_{0+} \tag{7.10e}$$

$$A_{km} \in \{0, 1\}\ \forall k, m \tag{7.10f}$$

$$\theta_k \in \mathcal{S}_k\ \forall k \in \mathcal{C}^{\mathrm{SC}}, \theta_k = 1\ \forall k \in \mathcal{C}^{\mathrm{MC}} \tag{7.10g}$$

In problem (7.10), constraints (7.10d) force each DP to be allocated to exactly one cell. Contraints (7.10c) represent a reformulation of the allocation rule introduced in Eq. (2.14). Problem (7.10) is a mixed-integer nonlinear problem (MINLP) because of the bilinear product terms $A_{km}\theta_k$. In the following, this problem is converted into a mixed integer linear problem (MILP) using the lifting strategy discussed in 3.2.2. Let the constant

$$\overline{\theta} = \underset{\tilde{s}, k}{\arg\max} \quad \delta_{\tilde{s},k} \tag{7.11}$$

denote the largest bias value. An auxiliary parameter Δ_{km} is introduced, for which $\Delta_{km} = A_{km}\theta_k\ \forall k, m$ is enforced using the following linear inequalities:

$$\Delta_{km} \leq A_{km}\overline{\theta} \tag{7.12a}$$

$$\Delta_{km} \leq \theta_{km} \tag{7.12b}$$

$$\Delta_{km} \geq \theta_{km} - (1 - A_{km})\overline{\theta} \tag{7.12c}$$

$$\Delta_{km} \geq 0 \tag{7.12d}$$

Problem (7.10) can be reformulated as the following:

$$\begin{aligned}
\underset{\Pi,\boldsymbol{A},\boldsymbol{\theta},\boldsymbol{\Delta}}{\text{minimize}} \quad & \Pi && \text{(7.13a)}\\
\text{subject to} \quad & \Pi \geq \sum_m A_{km}\Phi(k,m)\ \forall k && \text{(7.13b)}\\
& \sum_k \Delta_{km} p_k g_{km} \geq (\theta_j - \Delta_{jm}) p_j g_{jm}\ \forall j,m && \text{(7.13c)}\\
& (7.10\text{d}), (7.12)\ \forall k,m && \text{(7.13d)}\\
& \alpha \in \mathbb{R}_{0+} && \text{(7.13e)}\\
& A_{km} \in \{0,1\}\ \forall k,m && \text{(7.13f)}\\
& \theta_k \in \mathcal{S}_k\ \forall k \in \mathcal{C}^{\text{SC}}, \theta_k = 1\ \forall k \in \mathcal{C}^{\text{MC}} && \text{(7.13g)}\\
& \Delta_{km} \in \mathbb{R}_{0+} && \text{(7.13h)}
\end{aligned}$$

Problem (7.13) is linear in all optimization variables and therefore classifies as a MILP, which can be solved using conventional state-of-the art solvers. Even though problem (7.13) is capable of obtaining the optimal bias values, the network needs to gather all information about SINR-levels, user demands etc. centrally to solve the problem. In the following a learning-based decentralized approach is introduced.

Denote as $\boldsymbol{\theta}^*$ the optimal bias values for a given network scenario obtained by solving problem (7.13). A vector of class labels $\boldsymbol{y}$ is computed with its elements $y_k = \{s|\theta_k = \delta_s\}$. In the following suitable attributes are designed for each small cell that are being mapped to corresponding features to be used in the proposed classification scheme.
Denote the attribute $G^{\text{SC}}(k)$ which is determined as $G^{\text{SC}}(k) = 1$ if small cell k is deployed on the edge of the coverage areas between two macro cells, and $G^{\text{SC}}(k) = 0$ otherwise, which is illustrated in Fig. 4.4. Which of these roles a small cell fulfills is known to the operator from the network architecture.
For the second set of attributes, define the index set

$$\mathcal{M}_k^{\{\tilde{s}\}} = \{m|\delta_s p_k g_{km} \geq p_j g_{jm} \forall j \in \mathcal{C}^{\text{MC}}\} \tag{7.14}$$

of DPs connected to cell k if bias value $\delta_{\tilde{s}}$ is used, for which the expected load of cell k is computed as

$$G^{\text{LD}}(k,\tilde{s}) = \sum_{m \in \mathcal{M}_k^{\{\tilde{s}\}}} \frac{d_m}{R_{km}}. \tag{7.15}$$

Similarly the expected sum load is computed that DPs in the coverage area of SC m operating with bias δ_s cause to the first and second neighboring cell in the allocation

defined by $\boldsymbol{A}^0$:

$$G^{\mathrm{PSL}}(k,s) = \sum_{m \in \mathcal{M}_k^{\{s\}}} A^0_{\kappa_k^{\mathrm{P}} m} \Phi(\kappa_k^{\mathrm{P}}, m) \tag{7.16}$$

and

$$G^{\mathrm{SSL}}(k,s) = \sum_{m \in \mathcal{M}_k^{\{s\}}} A^0_{\kappa_k^{\mathrm{S}} m} \Phi(\kappa_k^{\mathrm{S}}, m), \tag{7.17}$$

respectively. The aforementioned attributes are combined into the following $(3S+1)$-element attribute vector:

$$\begin{aligned} \boldsymbol{h}_m = \Big[& G^{\mathrm{SC}}(k), G^{\mathrm{LD}}(k,1), \ldots, G^{\mathrm{LD}}(k,\tilde{S}), \\ & G^{\mathrm{PSL}}(k,1), \ldots, G^{\mathrm{PSL}}(k,\tilde{S}), G^{\mathrm{SSL}}(k,1), \ldots, G^{\mathrm{SSL}}(k,\tilde{S}) \Big]^\top \end{aligned} \tag{7.18}$$

This attribute vector is used to train SVMs that can classify pairwise between all available bias values, using the soft-margin introduced SVM in Sec. 3.3.2. The multi-class SVM training and classification problem is solved using the one-versus-one majority voting approach introduced in Sec. 3.3.3.

7.4 Simulation Results

Simulations of a wireless communication network with three macro- and six small cells in fixed positions as illustrated in Fig. 7.1 are carried out. The common network parameters from Table 4.4 are used.
Problem (2.13) is solved using the CVX toolbox for MATLAB [GB14] with the Gurobi solver [GUR], and the SVM training problem (3.23) is solved using the Machine Learning Toolbox for Matlab. For the training of the SVMs, 10000 DP attribute vectors are used from 100 simulations of network scenarios with $M = 100$ DPs each. The soft threshold weighting parameter C in problem (3.23) is determined by searching on a grid the value that provides the highest classification accuracy on the training set. For the function $\vartheta(\cdot)$ in problem (3.23) consider both, the linear mapping of attributes to features, and the mapping to quadratic features. In the following these two methods are referred to as "lin. SVM" and "quad. SVM", respectively. To evaluate the performance of the algorithms with a testing set, 100 instances of network scenarios are generated with $M = 100$ DPs each and the resulting load levels of each cell are averaged over all scenarios. Accordingly, SVM classification on the testing set is performed using the coefficients obtained from the training set.

As observable in Fig. 7.2, the maximum cell load increases with the demand, here simulated in the range of 0-1 MBit/s per DP. For the given network configuration, there are differences in the sizes of coverage areas that result in unbalanced load levels across the cells. The max. SINR approach is not designed to mitigate this imbalance, and therefore exhibits the worst performance. Both SVM-based methods however perform better than the max. SINR approach, with the quadratic SVM being very close to the optimal solution. This demonstrates that using the learning-based approach discussed in the paper, a decentralized load balancing scheme can be obtained that is close to the globally optimal solution of a computationally extensive, joint optimization of all allocations in the network.
For an increasing number of SCs randomly deployed in one of the locations shown in Fig. 7.1, the averaged maximum load also decreases linearly, as shown in Fig. 7.3. The average load level of individual cells for all methods and a fixed DP demand of 1 MBit/s is shown in Fig. 7.4. It shows that the cell MC1, which corresponds to the macro cell in the lower center area of Fig. 7.1, is close to being overloaded. Both the SVM-based methods and the optimal solution achieve this through offloading to small cell. It is observable that the small gap to the optimal solution probably originated from small cell SC2 being underutilized in the SVM-based methods compared to the optimal solution.

For the SVM training of the range expansion optimization scheme in Sec. 7.3, 250 network scenarios with nine small cells each are used for a total of 2250 training datapoints. To test the performance of the SVM-based classifier as a parameter optimization scheme, 100 new network scenarios are used in a Monte-Carlo evaluation and compute the average achieved cell loads as performance metrics.
As a benchmark to evaluate the performance of the proposed scheme, a network with SCs operating without range expansion and DP allocation according to the strongest received signal, in the following referred to as “no range exp.”, as defined by Eq. 2.14. The upper bound performance benchmark is given by the optimal bias selection obtained from solving problem (7.13). The averaged maximum load achieved by all schemes for range expansion over an increasing DP demand is illustrated in Fig. 7.6. It is observable that, similarly to the DP allocation optimization scheme, the scheme based on an SVM with a quadratic feature mapping performs close to optimal. The overall achievable load decrease however, even when solved optimally, is lower for the range expansion optimization approach. Fig.7.7 shows the average maximum load over the evaluated network scenarios for an increasing number of users with a data demand of 1 MBit/s each. As observable, the quadratic SVM achieves close to optimal performance, while the linear SVM causes slightly higher cell load, with both approaches showing lower load levels than without range expansion for all M. This underlines the

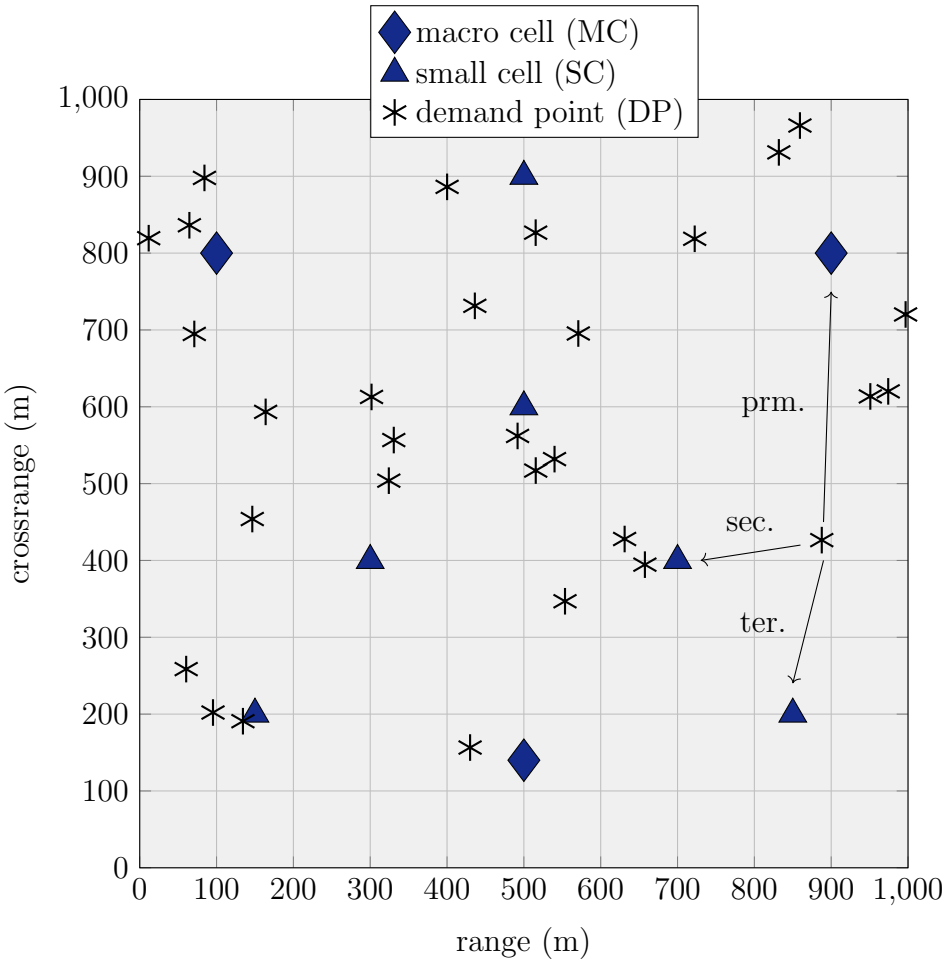

Figure 7.1. Illustration of the network scenario and primary, secondary and tertiary allocation candidates.

stability of the proposed scheme and the suitability of the selected SC features.
The load of individual cells for a simulation of 100 network scenarios with 100 DPs with 0.8 MBit/s data demand each is shown in Fig. 7.8. The load of MC1 without range expansion is the critical one to be minimized for the load balancing scheme to be successful. All proposed methods achieve decreased load for MC1, with the SVM-based approaches being only slightly worse than the optimum. The highest load of any SC is about 20%, which is a large increase relative to the load level without range expansion. The confusion matrix of optimal bias levels and classified bias levels for the quadratic SVM is shown in Fig. 7.9. The classifier shows very good performance with 93% accuracy in detecting which small cells, according to the optimal solution of the MILP, do not serve any DPs.Mainly for the bias values 0dB and 4dB, the accuracy is decreased. The most common error made by the classifier, with respect to the optimal MILP solution, is to not allocate users to SCs that for the optimal solution actually had users

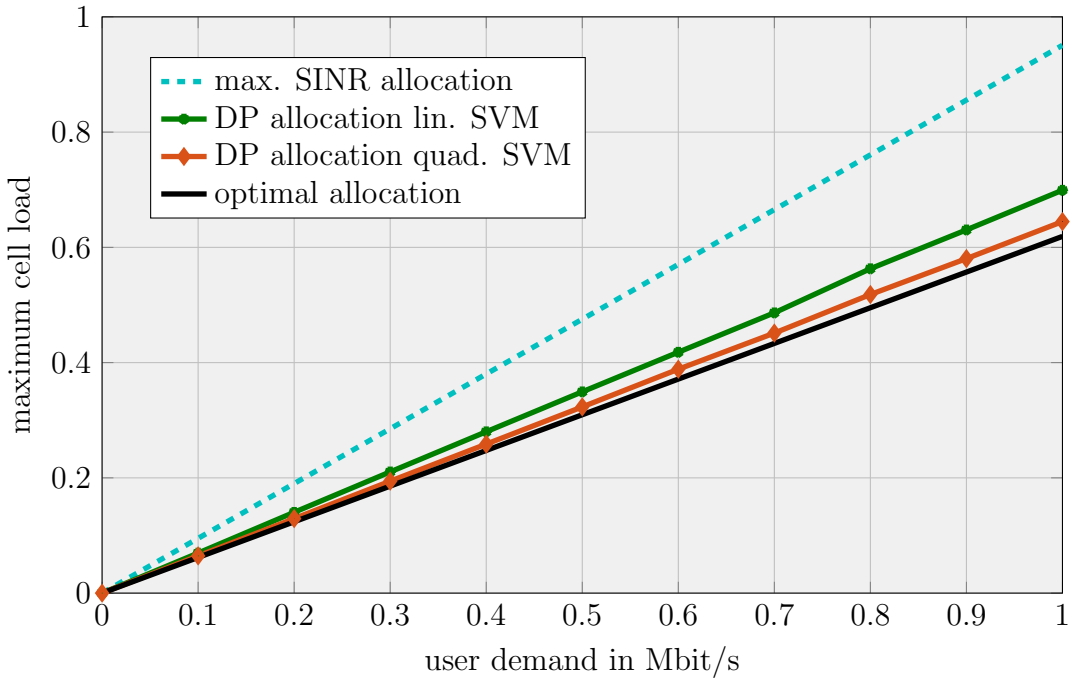

Figure 7.2. Maximum cell load comparison for learning-based and optimal user allocation over increasing demand. The SVM based on quadratic feature mapping performs close to optimal.

allocated to them. The good performance in load balancing however, as discussed for Fig. 7.7, suggests that these wrong classifications do not occur in critical scenarios.

7.5 Summary

A scheme was introduced to utilize multi-class SVMs for solving resource allocation ILPs arising in load balancing problems. The proposed method relies on training a classifier based on support vector machines using historical network data. This classifier is then used by each SC in operation of the network or each DP to make allocation decisions using locally available information, such that the maximum load of any cell in the network is minimized. Simulation results show that the proposed methods achieve close to optimal performance especially if support vector machines with quadratic feature mapping are being used.

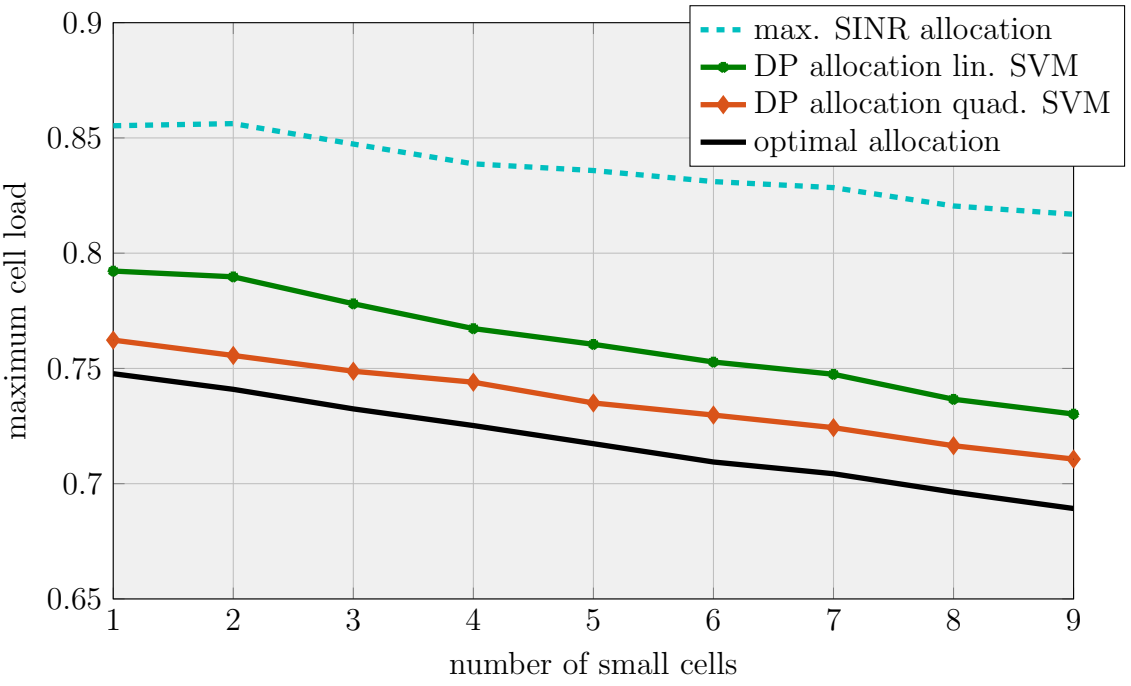

Figure 7.3. Maximum cell loads for user allocation schemes over number of deployed small cells. The deployment of additional SCs continuously decreases the maximum load level.

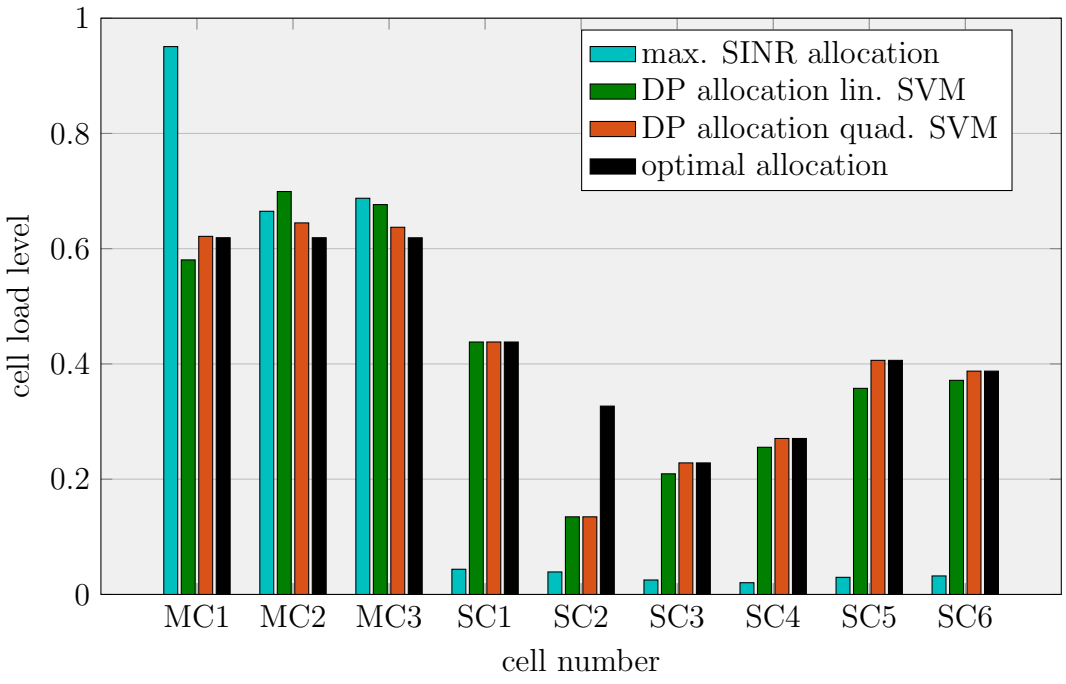

Figure 7.4. Example of cell loads for individual cells. All proposed approaches for SC allocation decrease the critical load level of MC1.

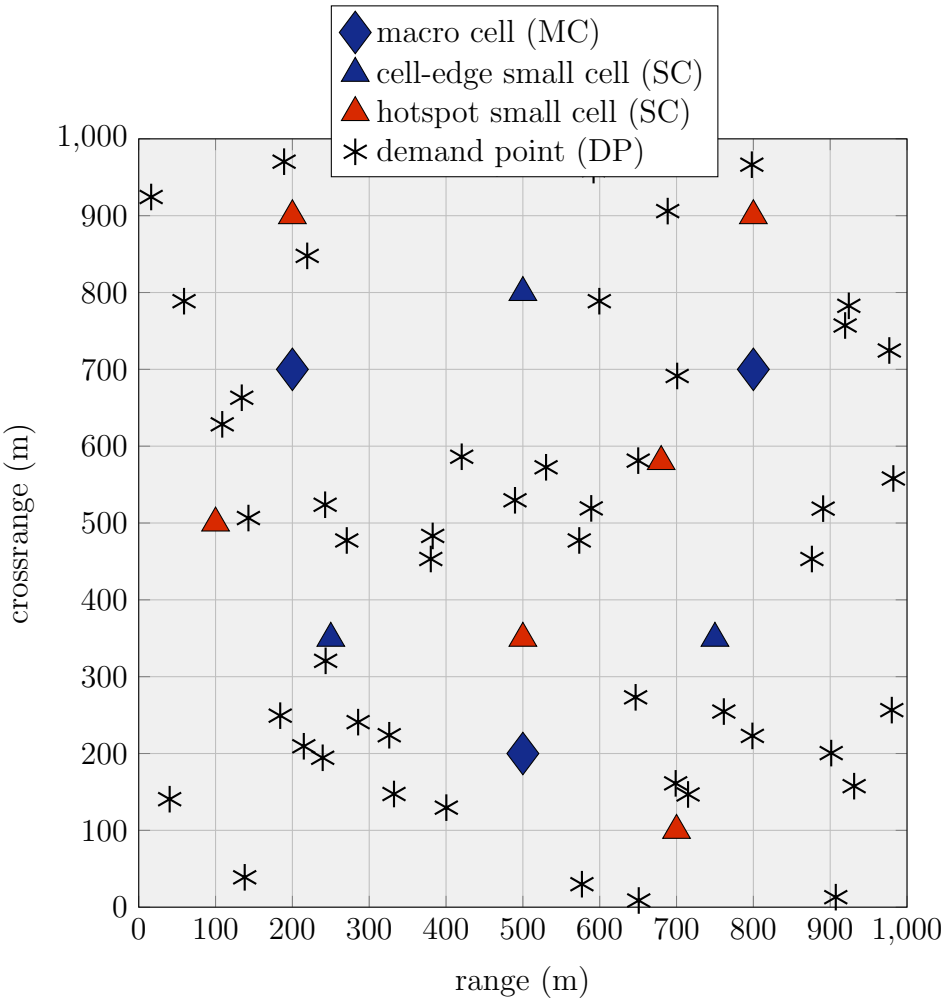

Figure 7.5. Illustration of the wireless network scenario and distinction between cell-edge versus hotspot SC types.

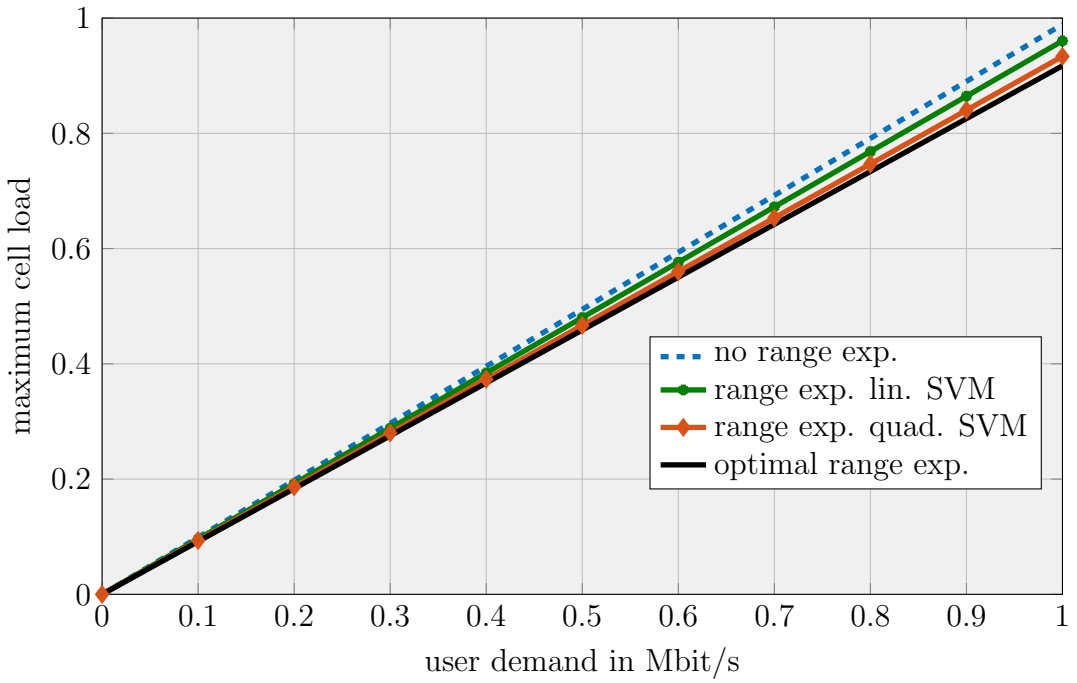

Figure 7.6. Maximum load level comparison for increasing user demand and different range expansion schemes. The load level increases linearly with the user demand.

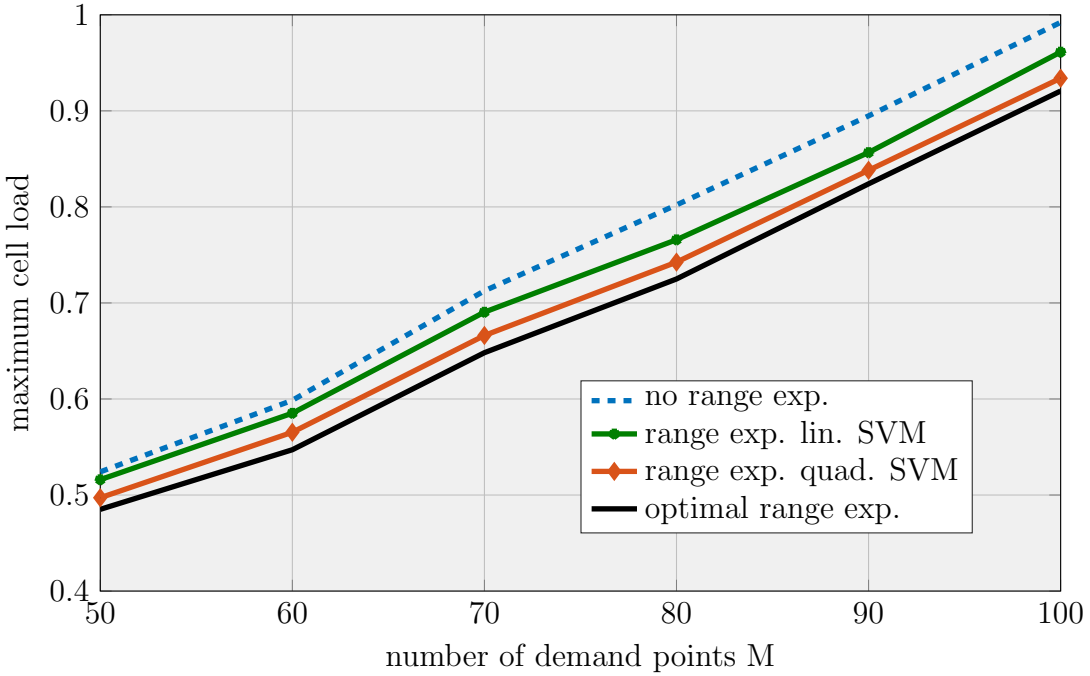

Figure 7.7. Maximum cell load levels for range expansion schemes over an increasing number of DPs.

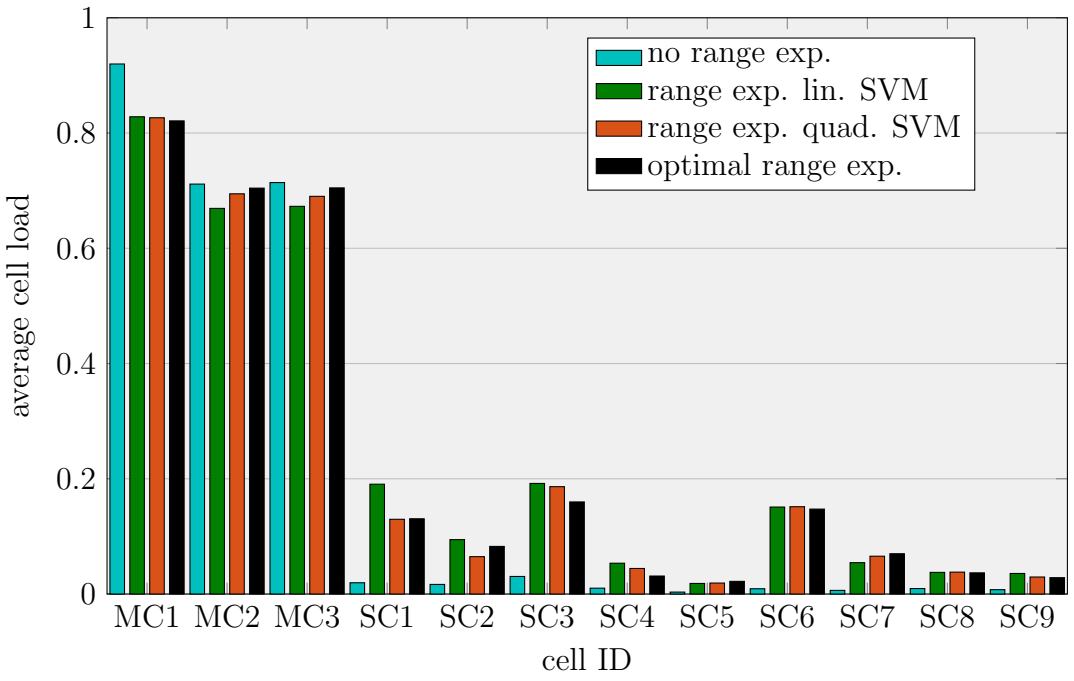

Figure 7.8. Cell load levels example for multiple range expansion schemes. The high number of small cells all remain underutilized without range expansion.

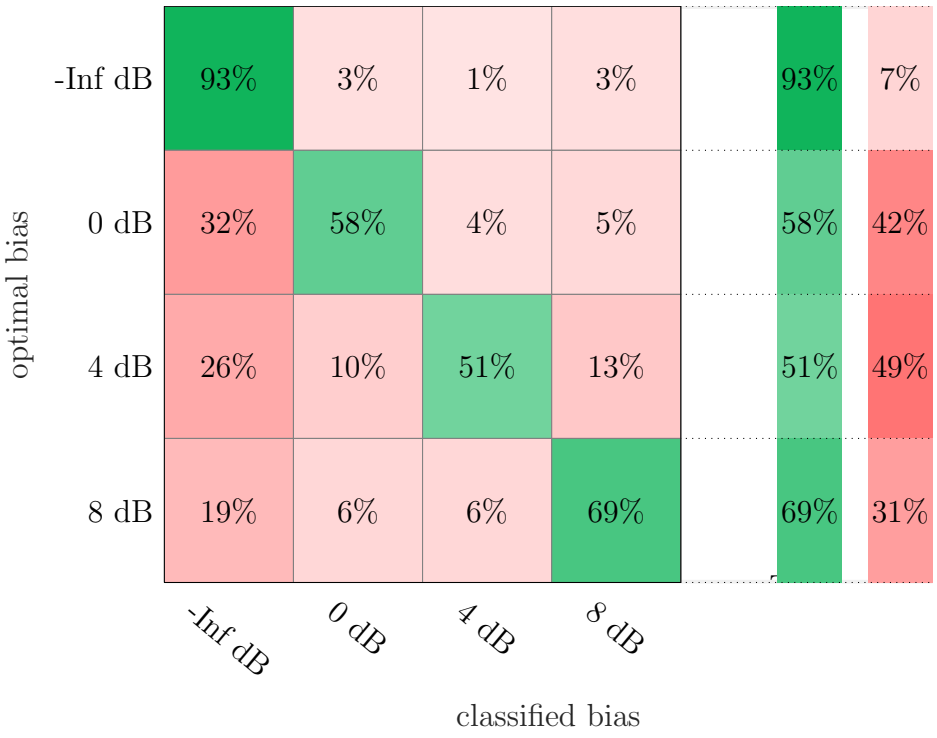

Figure 7.9. Confusion matrix for SVM with quadratic feature mapping.

Chapter 8

Conclusions and Outlook

In this thesis, a plan to enable gains in network performance for heterogeneous wireless communication networks is proposed. The performance of future wireless networks is measured in throughput, energy efficiency, spectral resource efficiency, adherence to QOS-constraints and other criteria. Similarly, any attempts to optimize the network with respect to these objectives can be applied on various timescales. Three optimization phases, in the order of long-term to more immediate measures, were identified: network deployment, network configuration and network operation. As the upcoming fifth and future generations of wireless networks need to provide a wide variety of services, all of which lead to QOS-constraints that need to be adequately addressed in the proposed approaches for network optimization. This favors a utilization of MIPs, for which reformulations to attain computational tractability are proposed as a major contribution of this thesis.

The problem of increasing the network performance, as measured by a variety of objectives, was divided into multiple sub-problems critical for this goal. Firstly, the network needs to be planned and set up to facilitate a load balanced operation. Only cells that are not overloaded or forced to operate at the expense of all their available resources, such as time-frequency resources or transmit power, can optimize their operation towards other objectives. Therefore, the network must be designed for load-balanced operation in the planning and scheduling phases. For the second objective, the network needs to be optimized towards an efficient utilization of the available resources. The third objective addresses economic operability considerations, where the energy consumption of all cells must be optimized in an effort to decouple the increase in network density from a proportional increase in energy consumption. Finally, the need of a load-balanced operation is again addresses by the fourth objective, which demands that load-balancing can be maintained in operation through fast and decentralized methods.

The first objective was addressed with a scheme to optimize the deployment location of SCs and their activity, specifically their on-off status over a given time horizon. The deployment location of SCs was optimized while considering area- and SC-type-dependent cost factors. As SCs in future HetNets are envisioned to operate with their own energy supply from renewable energies and utilizing energy storage, the optimization of their activity over a time horizon requires joint scheduling optimization

over multiple time instances. An optimization approach for grouping demand forecasts for multiple time instances into time-slots of varying length was proposed, based on the forecasted demand variability of the network. The cell activity was optimized based on the thus obtained time-schedule. Network simulations demonstrated the beneficial effect of the optimized deployment locations, activity status and scheduling timeframe for the load balancing of the network. The analysis shows that an optimization-based approach to planning the deployment locations of multiple cells jointly achieves lower cell load levels than a heuristic approach where the cells are deployed one-by-one. This has already been shown in [SY13], but the analysis in this work shows that the benefit is emphasized especially if multiple candidate locations and cells types are available. Furthermore, the joint optimization of cell activity and time schedule achieves lower load levels than optimizing the activity based on a schedule with timeslots of equal length, especially if there is high temporal variance in the spatial load distribution of the network.

The resource allocation of the HetNet was optimized in order to fulfill the second objective. The high variety of services provided by future HetNets necessitate viewing the network as the joint operation of multiple slices, which may utilize different time-frequency resources. A nonconvex MIP to jointly optimize the resource dimensioning of these slices, the allocation of cells to different slices, and the allocation of DPs to cells, was formulated. An inner linear approximation of the original problem was provided in the form of an MILP, that under certain conditions and with sufficient computational effort could solve the original problem optimally. It was demonstrated through simulation results that the proposed cell planning approach minimized the resource consumption of the network. Also it was demonstrated that the proposed approach, when operating with multiple orthogonal resource slices, enabled significant and reliable gains in resource efficiency through DP deployment. This result adresses the key challenge raised by the authors of [AZDG16], which is that novel network control mechanisms need to be developed for dense wireless networks to enable performance gains through network densification. The results in this work show that through a joint cell and spectrum planning approach, resource efficiency gains can be reliably achieved when additional small cells are deployed.

For the third objective of energy consumption minimization and economic operability, a scheme to minimize the total energy consumption of the network subject to QOS-constraints was proposed. The proposed MILP is an inner approximation of the original, computationally intractable power minimization problem. Other than established, heuristic approaches for power scaling in HetNets, multiple network parameters and QOS-constraints can be adequately modeled by relying on the solution of the MILP. A proof that the solution of the approximate MILP is always feasible for the original

MIP was also provided. Simulation results showed the superiority to established methods and demonstrated significant decreases in energy consumption. Most significantly, the proposed approach achieved lower energy consumption when compared with an exhaustive search scheme over all possible cell activity configurations combined with a heuristic power scaling approach introduced in [HYLS15]. The fundamental problem of economic operability regarding energy consumption of dense wireless networks raised in [CSS+14] and [AZDG16] can be effectively mitigated using the approach presented in this thesis.

Finally, the fourth objective of maintaining the load-balanced network state was addressed using two decentralized learning-based schemes. Communication and coordination overhead necessary for network-wide optimizations mitigates the feasibility of such schemes for a live application during network operation. A learning-based scheme was proposed that utilizes multi-class SVMs with locally available network attributes to perform decentralized load balancing. These SVMs, even though they are traditionally used as classifiers, were adapted to approximately solve network optimization IPs. Two approaches were proposed, one where DPs utilized the learning-based classifier to allocate to the best cells, and one where the SCs synthetically expanded their coverage area based on the learning system. Both approaches yielded the desired load-balancing effect, with almost the same performance as a global network optimization approach. In comparison with established load balancing approaches [SY12b, YRC+13], the proposed methods require only very limited local information exchange to achieve close-to-optimal performance. This enables their scalability to very large network scenarios.

Even though the developed approaches for network optimization effectively solved the problems defined at the beginning of the thesis, the resulting observations suggest some important follow-up research. The processing time, especially for the resource- and energy consumption optimization on a large network, remains very large even for the computationally tractable linear inner approximations. Possibly a close to optimal solution could be obtained by segmenting larger networks into separate clusters, applying the proposed schemes on each individually, and then fusing the results to obtain the global network optimization solution. Additionally, the currently significant popularity and proven performance of solution approaches based on deep learning strongly suggest more detailed research on possible applications to wireless network optimization.

List of Acronyms

4G	Fourth Generation Mobile Networks
5G	Fifth Generation Mobile Networks
AWGN	Additive White Gaussian Noise
CSI	Channel State Information
DP	Demand Point
EHF	Extremely High Frequency
eMBB	Enhanced Mobile Broadband
GSM	Global System for Mobile communications
HetNet	Heterogeneous Wireless Communication Network
ILP	Integer Linear Problem
IP	Integer Problem
ITU	International Telecommunications Union
LOS	Line-of-Sight
LTE	Long Term Evolution
LTE-A	Long Term Evolution Advanced
M2M	Machine-to-Machine
MC	Macro Cell
MILP	Mixed-Integer Linear Problem
MIMO	Multiple Input Multiple Output
MINLP	Mixed-Integer Nonlinear Problem
MIP	Mixed-Integer Problem
mMIMO	Massive MIMO
mMTC	Massive Machine Type Communications
mmWave	Millimeter-Wave

MRC Maximum Ratio Combining

NLOS Non-Line-of-Sight

NP Nondeterministic Polynomial Time

OFDMA Orthogonal Frequency Division Multiple Access

QOS Quality-of-Service

RAT Radio Access Technology

RF Radio Frequency

SC Small Cell

SINR Signal-to-Interference-Plus-Noise-Ratio

SDMA Space Division Multiple Access

SNR Signal-to-Noise-Ratio

SVM Support Vector Machine

UHF Ultra-High Frequency

URLLC Ultra-Reliable and Low-Latency Communications

V2V Vehicle-to-Vehicle

VR Virtual Reality

WLAN Wirless Local Area Network

ZF Zero Forcing

List of Symbols and Notation

Symbols and Functions:

A_{km}	binary allocation indicator of DP m to cell k
B_{qk}	binary allocation indicator of slice q to cell k
C	weighting factor of SVM soft-margin penalty term
d_m	data demand of DP m in bit per second
E_k	energy level of cell k in Joule
$E_k^{(\Gamma)}$	energy consumption of cell k based on model $\Gamma(\cdot)$
$F^{\text{TYPE..}}(\cdot)$	SVM attributes for decentralized user allocation
$G^{\text{SC..}}(\cdot)$	SVM attributes for decentralized SC range expansion
g_{km}	total link attenuation between cell k and DP m
h_{km}^{CH}	propagation channel coefficients between cell k and DP m
$\boldsymbol{H}$	attribute matrix for SVM training
$\boldsymbol{h}$	attribute vector of a single sample for SVM training
i	index of linearization functions, $i = 1, \ldots, I$
J_{st}	binary allocation indicator of snapshot s to time-slot t
k	index of cells, $k = 1, \ldots, K$
$L_n^{\text{P/S/R}}$	weighting factor of prim./sec./remaining interference in scenario n
l_t	length of time-slot t in seconds
m	index of DPs, $m = 1, \ldots, M$
n	index of interference scenarios, $n = 1, \ldots, N$
$\tilde{n}$	index of small cell models $\tilde{n} = 1, \ldots, \tilde{N}$
$P_k^{\text{MIN/MAX}}$	minimum/maximum transmit power of cell k in Watts
p_k	transmit power of cell k in Watts
q	index of network slices, $q = 1, \ldots, Q$
R_{km}	data rate achievable for cell k serving DP m in bits per second
$\tilde{s}$	index of available bias values for range expansion, $\tilde{s} = 1, \ldots, \tilde{S}$
s	index of snapshots, $s = 1, \ldots, S$
T_0	time constant for energy consumption model, in seconds
t	index of time-slots, $t = 1, \ldots, T$
$\tilde{t}$	index of attribute vectors for SVM, $\tilde{t} = 1, \ldots, \tilde{T}$
U_m	number of discrete users in DP m
$u_i(\cdot)$	piecewise linearizing function i
$v(\cdot)$	demand variability function

W	total availiable system bandwidth in Hz
w_q	system bandwidth allocated to slice q
$\overline{w}$	system bandwidth resources availiable for distribution between slices
$\boldsymbol{y}$	label vector for SVM training
Z	unused spectral resources in Hz
z_k	binary activity indicator of cell k
$\gamma_k m$	SINR of cell k serving DP m
$\Gamma(\cdot)$	model function for cell energy consumption
$\delta_{\tilde{s}}$	bias value with index $\tilde{s}$
ϵ	linearization accuracy parameter
$\zeta(\cdot)$	load term function, $\zeta(x) = 1/\log_2(1+x)$
η_{km}^{BW}	bandwidth efficiency of the link between cell k and DPm
θ_k	bias value of cell k used for cell range expansion
$\Theta_{\tilde{n}k}$	binary indicator of SC type $\tilde{n}$ deployment in cell location k
$\kappa_{km}^{\mathrm{P/S}}$	index of primary/secondary interferers for cell k serving DB m
λ	power spectral density of AWGN
ν_i	binary indicator of line segment i used for piecewise linearization
ξ	weighting factor for big-M method
Π	upper bound of cell loads minimized in load balancing
ρ_k	load of cell k
σ^2	signal power of AWGN
$\tau^{\mathrm{MIN/MAX}}$	minimum/maximum link load parameters
$\psi_{\tilde{t}}$	SVM misclassification penalty term of attribute vector $\tilde{t}$
Ψ_{nkm}	discrete interference scenario n for cell k serving DP m
$\varpi_{\tilde{n}}$	deployment cost of small cell model $\tilde{n}$
$\chi_{\tilde{n}/k}^{\mathrm{SC/LOC}}$	SC deployment cost factor for type $\tilde{n}$ / location k
$\Upsilon_{\tilde{n}km}$	binary offloading indicator for SC type $\tilde{n}$ in location k serving DP m

Sets:

$\emptyset$	the empty set
$\{0, 1\}$	set of binaries
$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set real numbers
$\mathbb{R}_{0+}$	set of nonnegative real numbers
$\mathcal{C}$	set of indices of all cells
$\mathcal{C}^{\mathrm{SC}}$	set of indices of small cells
$\mathcal{C}^{\mathrm{MC}}$	set of indices of macro cells
$\mathcal{M}$	set of indices of all DPs
$\mathcal{M}_k^{\{\tilde{s}\}}$	set of DPs in coverage area of SC k with bias $\tilde{s}$
$\mathcal{B}$	set of three binaries used for bilinear reformulation
$\mathcal{L}$	set of two bounded real scalars and a binary used for bilinear reformulation
$\mathcal{A}$	set of allocation parameters over a time horizon
$\mathcal{R}$	set of cell loads over a time horizon
$\mathcal{Z}$	set of cell activity indicators over a time horizon
$\mathcal{S}_k$	set of bias values available for cell k

Notation:

$\in$	element of
$\forall$	for all
$\subset$	is a proper subset of
$\cup$	set union
$\cap$	set intersection
A_{km}	element in the k-th row and m-th column of matrix $\boldsymbol{A}$
$\mathbb{R}^{K \times M}$	matrix with K rows and M columns of real parameters
$[\cdot]^\top$	vector transpose
$\mathbb{E}(\cdot)$	expected value
$\|\|\mathcal{M}\|\|$	number of elements in set $\mathcal{M}$
$\|\cdot\|$	magnitude

List of Figures

List of Tables

Bibliography

[3GP12] "3GPP Technical Report 36.839, V11.1.0 Mobility Enhancements in Heterogeneous Networks," Dec. 2012. [Online]. Available: https://portal.3gpp.org/

[3GP16] "3GPP Technical Report 38.913, Study on Scenarios and Requirements for Next Generation Access Technologies," Oct. 2016. [Online]. Available: https://portal.3gpp.org/

[ABC+14] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[ACM03] E. Amaldi, A. Capone, and F. Malucelli, "Planning UMTS Base Station Location: Optimization Models With Power Control and Algorithms," *Wireless Communications, IEEE Transactions on*, vol. 2, no. 5, pp. 939–952, Sept 2003.

[ADARCA17] A. Al-Dulaimi, S. Al-Rubaye, J. Cosmas, and A. Anpalagan, "Planning of Ultra-Dense Wireless Networks," *IEEE Network*, vol. 31, no. 2, pp. 90–96, March 2017.

[AFG04] W. P. Adams, R. J. Forrester, and F. W. Glover, "Comparisons and Enhancement Strategies for Linearizing Mixed 0-1 Quadratic Programs," *Discrete Optimization*, vol. 1, no. 2, pp. 99–120, 2004.

[ApS17] M. ApS, *The MOSEK Optimization Toolbox for MATLAB Manual. Version 8.1.*, 2017. [Online]. Available: http://docs.mosek.com/8.1/toolbox/index.html

[ARFB10] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power Consumption Modeling of Different Base Station Types in Heterogeneous Cellular Networks," in *2010 Future Network Mobile Summit*, June 2010, pp. 1–8.

[AWFT15] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing Client Association for Load Balancing and Fairness in Millimeter-Wave Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 23, no. 3, pp. 836–850, June 2015.

[AZDG16] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta, "Are we approaching the fundamental limits of wireless network densification?"

IEEE Communications Magazine, vol. 54, no. 10, pp. 184–190, October 2016.

[BC11] H. Bogucka and A. Conti, "Degrees of Freedom for Energy Savings in Practical Adaptive Wireless Systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 38–45, June 2011.

[BHL+14] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[BKD13] E. Björnson, M. Kountouris, and M. Debbah, "Massive MIMO and Small Cells: Improving Energy Efficiency by Optimal Soft-Cell Coordination," in *ICT 2013*, May 2013, pp. 1–5.

[BKL+13] P. Belotti, C. Kirches, S. Leyffer, J. Linderoth, J. Luedtke, and A. Mahajan, "Mixed-integer Nonlinear Optimization," *Acta Numerica*, vol. 22, p. 1131, 2013.

[BL12] S. Burer and A. N. Letchford, "Non-Convex mixed-Integer Nonlinear Programming: A Survey," *Surveys in Operations Research and Management Science*, vol. 17, no. 2, pp. 97–106, 2012.

[BLM+14] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, February 2014.

[CBD11] D.-S. Chen, R. G. Batson, and Y. Dang, *Applied Integer Programming: Modeling and Solution*. John Wiley & Sons, 2011.

[CBdVCP17] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Prez, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Transactions on Networking*, vol. 25, no. 5, pp. 3044–3058, Oct 2017.

[CG17] H. Celebi and . Gven, "Load Analysis and Sleep Mode Optimization for Energy-Efficient 5G Small Cell Networks," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2017, pp. 1159–1164.

[Cim85] L. Cimini, "Analysis and Simulation of a Digital Mobile Channel Using Orthogonal Frequency Division Multiplexing," *IEEE Transactions on Communications*, vol. 33, no. 7, pp. 665–675, jul 1985.

[Cis17] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016-2021," Sep. 2017. [Online]. Available: https://www.cisco.com/

[CL11] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[CPP13] Y. Cheng, M. Pesavento, and A. Philipp, "Joint Network Optimization and Downlink Beamforming for CoMP Transmissions Using Mixed Integer Conic Programming," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 3972–3987, Aug 2013.

[CS02] K. Crammer and Y. Singer, "On the Learnability and Design of Output Codes for Multiclass Problems," *Machine Learning*, vol. 47, no. 2-3, pp. 201–233, 2002.

[CSS+14] R. Cavalcante, S. Stanczak, M. Schubert, A. Eisenblaetter, and U. Tuerke, "Toward Energy-Efficient 5G Wireless Communications Technologies: Tools for Decoupling the Scaling of Networks from the Growth of Operating Power," *Signal Processing Magazine, IEEE*, vol. 31, no. 6, pp. 24–34, Nov 2014.

[CV95] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[CZB+10] L. M. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. Gdor, G. Auer, and L. V. D. Perre, "Challenges and Enabling Technologies for Energy Aware Mobile Radio Networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 66–72, November 2010.

[CZL16] J. Chen, H. Zhuang, and Z. Luo, "Energy Optimization in Dense OFDM Networks," *IEEE Communications Letters*, vol. 20, no. 1, pp. 189–192, Jan 2016.

[Dak65] R. J. Dakin, "A Tree-Search Algorithm for Mixed Integer Programming Problems," *The Computer Journal*, vol. 8, no. 3, pp. 250–255, 1965. [Online]. Available: http://dx.doi.org/10.1093/comjnl/8.3.250

[DDG+12] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. J. Gonzalez, H. Klessig, I. Gdor, M. Olsson, M. A. Imran, A. Ambrosy, and O. Blume, "Flexible Power Modeling of LTE Base Stations," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2012, pp. 2858–2862.

[DJM14] M. Deruyck, W. Joseph, and L. Martens, "Power Consumption Model for Macrocell and Microcell Base Stations," *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 3, pp. 320–333, 2014. [Online]. Available: http://dx.doi.org/10.1002/ett.2565

[DMW+11] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A Survey on 3GPP Heterogeneous Networks," *Wireless Communications, IEEE*, vol. 18, no. 3, pp. 10–21, June 2011.

[DYFP14] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the Performance of Non-Orthogonal Multiple Access in 5G Systems with Randomly Deployed Users," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1501–1505, Dec 2014.

[EHF08] S. Elayoubi, O. B. Haddada, and B. Fourestie, "Performance Evaluation of Frequency Planning Schemes in OFDMA-Based Networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1623–1633, May 2008.

[FAC89] C. A. Floudas, A. Aggarwal, and A. R. Ciric, "Global Optimum Search for Nonconvex NLP and MINLP Problems," *Computers & Chemical Engineering*, vol. 13, no. 10, pp. 1117 – 1132, 1989. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0098135489870164

[Fet14] G. P. Fettweis, "The Tactile Internet: Applications and Challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, March 2014.

[FL05] C. A. Floudas and X. Lin, "Mixed Integer Linear Programming in Process Scheduling: Modeling, Algorithms, and Applications," *Annals of Operations Research*, vol. 139, no. 1, pp. 131–162, Oct 2005. [Online]. Available: https://doi.org/10.1007/s10479-005-3446-x

[FMK17] X. Foukas, M. K. Marina, and K. Kontovasilis, "Orion: RAN Slicing for a Flexible and Cost-Effective Multi-Service Mobile Network Architecture," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017, pp. 127–140.

[FSPRA18] R. Ferrs, O. Sallent, J. Prez-Romero, and R. Agusti, "On the Automation of RAN Slicing Provisioning and Cell Planning in NG-RAN," in *2018 European Conference on Networks and Communications (EuCNC)*, June 2018, pp. 37–42.

[FWL+17] W. Feng, Y. Wang, D. Lin, N. Ge, J. Lu, and S. Li, "When mmWave Communications Meet Network Densification: A Scalable Interference Coordination Perspective," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1459–1471, July 2017.

[GACD13] A. Gupte, S. Ahmed, M. S. Cheon, and S. Dey, "Solving Mixed Integer Bilinear Problems Using MILP Formulations," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 721–744, 2013.

[GB08] M. C. Grant and S. P. Boyd, "Graph Implementations for Nonsmooth Convex Programs," in *Recent Advances in Learning and Control*. Springer, 2008, pp. 95–110.

[GB14] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[GJ15] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[GKN+15] S. Gulati, S. Kalyanasundaram, P. Nashine, B. Natarajan, R. Agrawal, and A. Bedekar, "Performance analysis of distributed multi-cell coordinated scheduler," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, Sept 2015, pp. 1–5.

[Gol04] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2004.

[Gom58] R. E. Gomory, "Outline of an Algorithm for Integer Solutions to Linear Programs," *Bulletin of the American Mathematical society*, vol. 64, no. 5, pp. 275–278, 1958.

[GTC+14] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. R. MacCartney, S. Sun, and S. Nie, "Millimeter-Wave Enhanced Local Area Systems: A High-Data-Rate Approach for Future Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1152–1163, June 2014.

[GTM+16] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G Ultra-Dense Cellular Networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72–79, February 2016.

[GUR] "Gurobi Optimizer 6.0," www.gurobi.com.

[HBB11] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green Cellular Networks: A Survey, Some Research Issues and Challenges," *IEEE Communications Surveys Tutorials*, vol. 13, no. 4, pp. 524–540, 4 2011.

[HBCO12] H. Hijazi, P. Bonami, G. Cornuéjols, and A. Ouorou, "Mixed-Integer Nonlinear Programs Featuring On/Off Constraints," *Computational Optimization and Applications*, vol. 52, no. 2, pp. 537–558, 2012.

[HIXR15] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-Scale Antenna Systems with Hybrid Analog and Digital Beamforming for Millimeter Wave 5G," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 186–194, January 2015.

[HKD11] J. Hoydis, M. Kobayashi, and M. Debbah, "Green Small-Cell Networks," *IEEE Vehicular Technology Magazine*, vol. 6, no. 1, pp. 37–43, March 2011.

[HL02] C.-W. Hsu and C.-J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar 2002.

[HLQ+14] R. W. Heath, G. Laus, T. Q. S. Quek, S. Talwar, and P. Zhou, "Signal Processing for the 5G Revolution [From the Guest Editors]," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 12–13, Nov 2014.

[HQ14] R. Q. Hu and Y. Qian, "An Energy Efficient and Spectrum Efficient Wireless Heterogeneous Network Framework for 5G Systems," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 94–101, May 2014.

[HRTA14] E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution Toward 5G Multi-Tier Cellular Wireless Networks: An Interference Management Perspective," *IEEE Wireless Communications*, vol. 21, no. 3, pp. 118–127, June 2014.

[HYLS15] C. K. Ho, D. Yuan, L. Lei, and S. Sun, "Power and Load Coupling in Cellular Networks for Energy Optimization," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 509–519, Jan 2015.

[IRH+14] C. L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward Green and Soft: A 5G Perspective," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, February 2014.

[itu15] "ITU-R, IMT Vision Framework and overall objectives of the future development of IMT for 2020 and beyond, Recommendation ITU-R

M.2083-0," Sep. 2015. [Online]. Available: https://www.itu.int/rec/R-REC-M.2083

[ITU17] ITU, "Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)," Nov. 2017. [Online]. Available: https://www.itu.int/pub/R-REP-M.2410-2017

[Iwa15] M. Iwamura, "NGMN View on 5G Architecture," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[JMZ+14] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfrojd, and T. Svensson, "The Role of Small Cells, Coordinated Multipoint, and Massive MIMO in 5G," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 44–51, May 2014.

[JZR+17] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. C. Chen, and L. Hanzo, "Machine Learning Paradigms for Next-Generation Wireless Networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, April 2017.

[Kar72] R. M. Karp, "Reducibility Among Combinatorial Problems," in *Complexity of computer computations*. Springer, 1972, pp. 85–103.

[KB02] S. Kandukuri and S. Boyd, "Optimal Power Control in Interference-Limited Fading Wireless Channels with Outage-Probability Specifications," *IEEE Transactions on Wireless Communications*, vol. 1, no. 1, pp. 46–55, Jan 2002.

[KBTV10] A. Khandekar, A. Bhushan, J. Tingfang, and V. Vanghi, "LTE-Advanced: Heterogeneous Networks," in *Wireless Conference (EW), 2010 European*, April 2010, pp. 978–982.

[KMK12] S. Kaneko, T. Matsunaka, and Y. Kishi, "A Cell-Planning Model for HetNet with CRE and TDM-ICIC in LTE-Advanced," in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, May 2012, pp. 1–5.

[KN13] E. Klotz and A. M. Newman, "Practical Guidelines for Solving Difficult Mixed Integer Linear Programs," *Surveys in Operations Research and Management Science*, vol. 18, no. 1-2, pp. 18–32, 2013.

[Kre99] U. H.-G. Kressel, "Pairwise Classification and Support Vector Machines," *Advances in Kernel Methods*, pp. 255–268, 1999. [Online]. Available: http://dl.acm.org/citation.cfm?id=299094.299108

[KU16] Q. Kuang and W. Utschick, "Energy Management in Heterogeneous Networks With Cell Activation, User Association, and Interference Coordination," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3868–3879, June 2016.

[LCG+13] M.-H. Lin, J. G. Carlsson, D. Ge, J. Shi, and J.-F. Tsai, "A Review of Piecewise Linearization Methods," *Mathematical Problems in Engineering*, 2013.

[LETM14] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[LKB+14] C. Lange, D. Kosiankowski, A. Betker, H. Simon, N. Bayer, D. von Hugo, H. Lehmann, and A. Gladisch, "Energy Efficiency of Load-Adaptively Operated Telecommunication Networks," *Journal of Lightwave Technology*, vol. 32, no. 4, pp. 571–590, Feb 2014.

[LPGdlR+11] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Quek, and J. Zhang, "Enhanced Intercell Interference Coordination Challenges In Heterogeneous Networks," *Wireless Communications, IEEE*, vol. 18, no. 3, pp. 22–30, June 2011.

[LS99] J. T. Linderoth and M. W. P. Savelsbergh, "A Computational Study of Search Strategies for Mixed Integer Programming," *INFORMS Journal on Computing*, vol. 11, no. 2, pp. 173–187, 1999. [Online]. Available: https://doi.org/10.1287/ijoc.11.2.173

[LSB+16] G. Lee, W. Saad, M. Bennis, A. Mehbodniya, and F. Adachi, "Online Ski Rental for Scheduling Self-Powered, Energy Harvesting Small Base Stations," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[LT15] M.-H. Lin and J.-F. Tsai, "Comparisons of Break Points Selection Strategies for Piecewise Linear Approximation," *International Journal of Mechanical Engineering and Robotics Research*, vol. 4, no. 3, pp. 247–250, July 2015.

[LYHS15] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Optimal Cell Clustering and Activation for Energy Saving in Load-Coupled Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6150–6163, Nov 2015.

[MAT] "MATLAB and Statistics Toolbox Release 2013a, The MathWorks, inc., Natick, Massachusetts, United States."

[MB09] A. Magnani and S. P. Boyd, "Convex Piecewise-Linear Fitting," *Optimization and Engineering*, vol. 10, no. 1, pp. 1–17, 2009.

[MCB09] A. Mitsos, B. Chachuat, and P. I. Barton, "McCormick-Based Relaxations of Algorithms," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 573–601, 2009.

[McC76] G. P. McCormick, "Computability of Global Solutions to Factorable Nonconvex Programs: Part I — Convex Underestimating Problems," *Mathematical Programming*, vol. 10, no. 1, pp. 147–175, Dec 1976. [Online]. Available: https://doi.org/10.1007/BF01580665

[MCLG06] R. Madan, S. Cui, S. Lall, and N. A. Goldsmith, "Cross-Layer Design for Lifetime Maximization in Interference-Limited Wireless Sensor Networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 11, pp. 3142–3152, November 2006.

[MGRD17] M. Miozzo, L. Giupponi, M. Rossi, and P. Dini, "Switch-on-off policies for energy harvesting small cells through distributed q-learning," in *2017 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, March 2017, pp. 1–6.

[MHLT11] G. Miao, N. Himayat, G. Y. Li, and S. Talwar, "Distributed Interference-Aware Energy-Efficient Power Optimization," *IEEE Transactions on Wireless Communications*, vol. 10, no. 4, pp. 1323–1333, April 2011.

[MHV+12] H.-L. Määttänen, K. Hämäläinen, J. Venäläinen, K. Schober, M. Enescu, and M. Valkama, "System-level performance of lte-advanced with joint transmission and dynamic point selection schemes," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 247, Nov 2012. [Online]. Available: http://dx.doi.org/10.1186/1687-6180-2012-247

[MK10] K. Majewski and M. Koonert, "Conservative Cell Load Approximation for Radio Networks with Shannon Channels and its Application to LTE Network Planning," in *2010 Sixth Advanced International Conference on Telecommunications (AICT)*, May 2010, pp. 219–225.

[MMR+01] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, Mar 2001.

[MMWW02] H. Marchand, A. Martin, R. Weismantel, and L. Wolsey, "Cutting Planes in Integer and Mixed Integer Programming," *Discrete Applied Mathematics*, vol. 123, no. 1, pp. 397 – 446, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166218X01003481

[MNK+07] P. Mogensen, W. Na, I. Kovacs, F. Frederiksen, A. Pokhariyal, K. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE Capacity Compared to the Shannon Bound," in *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, April 2007, pp. 1234–1238.

[NGM15] "NGMN Alliance: NGMN 5G White Paper," Mar. 2015. [Online]. Available: https://www.ngmn.org/5g-white-paper/5g-white-paper.html

[NH09] D. Niyato and E. Hossain, "Dynamics of Network Selection in Heterogeneous Wireless Networks: An Evolutionary Game Approach," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 2008–2017, May 2009.

[NK17] V. M. Nguyen and M. Kountouris, "Performance Limits of Network Densification," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1294–1308, June 2017.

[NKDA18] Q. U. A. Nadeem, A. Kammoun, M. Debbah, and M. S. Alouini, "Design of 5G Full Dimension Massive MIMO Systems," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 726–740, Feb 2018.

[NLM13] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems," *IEEE Transactions on Communications*, vol. 61, no. 4, pp. 1436–1449, April 2013.

[OBB+14] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014.

[P1 16] *Description of Network Slicing Concept*, P1 WS1 E2E Architecture team Std., Rev. 1.0.8, 9 2016. [Online]. Available: http://www.ngmn.org/publications/all-downloads/article/update-to-ngmn-description-of-network-slicing-concept.html

[PCST00] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAGs for Multiclass Classification," in *Advances in neural information processing systems*, 2000, pp. 547–553.

[RCBHP17a] O. D. Ramos-Cantor, J. Belschner, G. Hegde, and M. Pesavento, "Centralized coordinated scheduling in LTE-Advanced networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, p. 122, Jul 2017. [Online]. Available: https://doi.org/10.1186/s13638-017-0904-5

[RCBHP17b] ——, "Centralized Coordinated Scheduling in LTE-Advanced Networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, p. 122, 2017.

[RMSS15] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband Millimeter-Wave Propagation Measurements and Channel Models for Future Wireless Communication System Design," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3029–3056, Sept 2015.

[RRE14] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.

[RSM+13] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[RSP+14] W. Roh, J. Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 106–113, February 2014.

[SAD+16] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-Enabled Tactile Internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, March 2016.

[SBSLa14] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-aho, "Opportunistic sleep mode strategies in wireless small cell networks," in *2014 IEEE International Conference on Communications (ICC)*, June 2014, pp. 2707–2712.

[Sch98] A. Schrijver, *Theory of Linear and Integer Programming.* John Wiley & Sons, 1998.

[SKYK11] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base Station Operation and User Association Mechanisms for Energy-Delay Tradeoffs in Green Cellular Networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1525–1536, September 2011.

[SMS+17] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. D. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[SP96] G. Schilling and C. Pantelides, "A Simple Continuous-Time Process Scheduling Formulation and a Novel Solution Algorithm," *Computers & Chemical Engineering*, vol. 20, pp. S1221 – S1226, 1996, european Symposium on Computer Aided Process Engineering-6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0098135496002116

[SPRFA17] O. Sallent, J. Perez-Romero, R. Ferrus, and R. Agusti, "On Radio Access Network Slicing from a Radio Resource Management Perspective," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 166–174, October 2017.

[SY12a] I. Siomina and D. Yuan, "Analysis of Cell Load Coupling for LTE Network Planning and Optimization," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2287–2297, June 2012.

[SY12b] ——, "Load Balancing in Heterogeneous LTE: Range Optimization via Cell Offset and Load-Coupling Characterization," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 1357–1361.

[SY13] ——, "Optimization Approaches for Planning Small Cell Locations in Load-Coupled Heterogeneous LTE Networks," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, Sept 2013, pp. 2904–2908.

[TG14] F. Trespalacios and I. E. Grossmann, "Review of Mixed-Integer Nonlinear and Generalized Disjunctive Programming Methods," *Chemie Ingenieur Technik*, vol. 86, no. 7, pp. 991–1012, 2014.

[TUY14] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-To-Device Communication in 5G Cellular Networks: Challenges, Solutions, and

Future Directions," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 86–92, May 2014.

[TV05] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge University Press, 2005.

[VHD+11] W. Vereecken, W. V. Heddeghem, M. Deruyck, B. Puype, B. Lannoo, W. Joseph, D. Colle, L. Martens, and P. Demeester, "Power Consumption in Telecommunication Networks: Overview and Reduction Strategies," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 62–69, June 2011.

[WCLM99] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, "Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[WHG+14] C. X. Wang, F. Haider, X. Gao, X. H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, February 2014.

[Wu97] T.-H. Wu, "A Note on a Global Approach for General 0–1 Fractional Programming," *European Journal of Operational Research*, vol. 101, no. 1, pp. 220–223, 1997.

[WWH+17] L. Wang, K. K. Wong, R. W. Heath, J. Yuan, and J. Yuan, "Wireless Powered Dense Cellular Networks: How Many Small Cells Do We Need?" *IEEE Journal on Selected Areas in Communications*, vol. PP, no. 99, pp. 1–1, 2017.

[XMH+17] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. Karagiannidis, E. Bjrnson, K. Yang, C.-L. I, and A. Ghosh, "Millimeter Wave Communications for Future Mobile Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1909–1935, Sept 2017.

[YGGY13] D. Yue, G. Guillén-Gosálbez, and F. You, "Global Optimization of Large-Scale Mixed-Integer Linear Fractional Programming Problems: A Reformulation-Linearization Method and Process Scheduling Applications," *AIChE Journal*, vol. 59, no. 11, pp. 4255–4272, 2013.

[YLY16] L. You, L. Lei, and D. Yuan, "Optimizing Power and User Association for Energy Saving In Load-Coupled Cooperative LTE," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[YPC+15] Z. H. Yang, Y. J. Pan, M. Chen, H. Xu, and J. F. Shi, "Cell Load Coupling With Power Control for LTE Network Planning," in *2015 International Conference on Wireless Communications Signal Processing (WCSP)*, Oct 2015, pp. 1–5.

[YRC+13] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[YY17] L. You and D. Yuan, "Load Optimization With User Association in Cooperative and Load-Coupled LTE Networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3218–3231, May 2017.

[ZHS10] M. Zulhasnine, C. Huang, and A. Srinivasan, "Efficient Resource Allocation for Device-to-Device Communication Underlaying LTE Network," in *2010 IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications*, Oct 2010, pp. 368–375.

[ZJ15] A. Zappone and E. Jorswieck, "Energy Efficiency in Wireless Networks via Fractional Programming Theory," *Foundations and Trends in Communications and Information Theory*, vol. 11, no. 3-4, pp. 185–396, 2015. [Online]. Available: http://dx.doi.org/10.1561/0100000088

[ZLC+17] H. Zhang, N. Liu, X. Chu, K. Long, A. H. Aghvami, and V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, 2017.

[ZSB+16] A. Zappone, L. Sanguinetti, G. Bacci, E. Jorswieck, and M. Debbah, "Energy-Efficient Power Control: A Look at 5G Wireless Technologies," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1668–1683, April 2016.

List of Publications

- Bahlke, F.; Ramos-Cantor, O.D.; Henneberger, S.; Pesavento, M.: *Optimized Cell Planning for Network Slicing in Heterogeneous Wireless Communication Networks*, IEEE Communication Letters 2018, Vol. 22 (8), pp. 1676-1679

- Bahlke, F.; Pesavento, M.: *Energy Consumption Optimization in Mobile Communication Networks*, submitted for journal publication (preprint: https://arxiv.org/abs/1807.02651)

- Bahlke, F.; Yang, J.; Pesavento, M.: *Activity Scheduling for Energy Harvesting Small Cells in 5G Wireless Communication Networks*, accepted for publication in the Proceedings of the 29th IEEE Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC 2018), September 2018

- Bahlke, F.; Pesavento, M.: *Optimized Small Cell Range Expansion in Mobile Communication Networks Using Multi-Class Support Vector Machines*, accepted for publication in the Proceedings of the 29th European Signal Processing Conference (EUSIPCO 2018), September 2018

- Bahlke, F.; Pesavento, M.: *Decentralized Load Balancing in Mobile Communication Networks*, Proceedings of the 25th IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP 2018), April 2018, pp. 3564-3568

- Bahlke, F.; Liu, Y.; Pesavento, M.: *Stochastic Load Scheduling for Risk Limiting Economic Dispatch in Smart Microgrids*, Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP 2016), March 2016, pp. 2479-2483

- Bahlke, F.; Ramos-Cantor, O.D.; Pesavento, M.: *Budget Constrained Small Cell Deployment Planning for Heterogeneous LTE Networks*, Proceedings of the 16th IEEE Workshop on Signal Processing Advances in Wireless Communications (IEEE SPAWC), June 2015, pp. 1-5

- Bahlke, F.; Zemmari, R.; Nickel, U.; Pesavento, M.: *Mismatch Loss Constrained Instrumental Variable Filtering for GSM Passive Bistatic Radar*, Proceedings of the Eighth IEEE Sensor Array and Multichannel Signal Processing Workshop (IEEE SAM 2014), June 2014, pp. 313-316

Curriculum Vitae

Name:	Florian Bahlke
Geburtsdatum:	17.08.1987
Geburtsort:	Offenbach

Schulausbildung

08/98 - 06/07	Franziskanergymnasium Kreuzburg, Großkrotzenburg, Schulabschluss: Allgemeine Hochschulreife

Studium

seit 02/14	Promotion in Elektrotechnik und Informationstechnik, Fachgebiet Nachrichtentechnische Systeme, Institut für Nachrichtentechnik, Technische Universität Darmstadt
07/11 - 12/13	Studium der Elektrotechnik und Informationstechnik, Technische Universität Darmstadt, Studienabschluss: Master of Science
10/07 - 07/11	Studium der Elektrotechnik und Informationstechnik, Technische Universität Darmstadt, Studienabschluss: Bachelor of Science

Berufstätigkeit

seit 02/14	Wissenschaftlicher Mitarbeiter, Fachgebiet Nachrichtentechnische Systeme, Institut für Nachrichtentechnik, Technische Universität Darmstadt
03/13 - 12/13	Wissenschaftlicher Mitarbeiter, Abteilung Sensordaten- und Informationsfusion, Fraunhofer FKIE, Wachtberg

Erklärung laut §9 der Promotionsordnung

Ich versichere hiermit, dass ich die vorliegende Dissertation allein und nur unter Verwendung der angegebenen Literatur verfasst habe. Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 29. Oktober 2018